ARI Research Note 98-09

The Role of Data and Feedback Error in Inference and Prediction

Michael Doherty, Ryan Tweney, Lowell Schipper and Raymond O'Connor
Bowling Green State University

Research and Advanced Concepts Office
Michael Drillings, Chief

June 1998

U.S. Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute for the Behavioral and Social Sciences

A Directorate of the U.S. Total Army Personnel Command

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

Bowling Green State University

Technical Review by

Michael Drillings, ARI

NOTICES

DISTRIBUTION: This Research Note has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This Research Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this Research Note are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE June 98	3. REPORT TYPE AND DATES COVERED Final Report 5 May 1985

4. TITLE AND SUBTITLE The Role of Data Feedback Error in Inference and Prediction (and

5. FUNDING NUMBERS MDA903-85-K-0193
PE 0601102A
20161102B74F
TA 1012
WU C06

6. AUTHOR(S) Michael Doherty, Ryan Tweney, Lowell Schipper & Raymond O'Connor

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Bowling Green State University , 120 Mcfall Center, Research SE, Bowling Green, OH 434038

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Army Research Institute, 5001 Eisenhower Ave, Alexandria, VA 22333-5600

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
Research Note 98-09

11. SUPPLEMENTARY NOTES COR: Michael Drillings

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)* The present research investigates two forms of uncertainty, defined operationally as error in the data, at two places within the information flow between the person and the environment. The two kinds were "measurement error" and "system failure error". The former involved adding a random variable to the data. The latter involved the sort of error which occurs when an environmental source of data gives information unrelated to the causal or predictive process under study. These forms of error were studied in the data and also in the feedback of subjects.

14. SUBJECT TERMS Measurement Error System Failure Error Information System

15. NUMBER OF PAGES 212

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

THE ROLE OF DATA AND FEEDBACK ERROR

IN

INFERENCE AND PREDICTION

Michael Doherty
Ryan Tweney
Lowell Schipper
Raymond O'Connor

Bowling Green State University
120 Mefall Center
Research SE
Bowling Green, OH 43408

FINAL REPORT
ARI Contract # MDA 903-85-K-0193
COR: Michael Drillings

THE ROLE OF DATA AND FEEDBACK IN INFERENCE AND PREDICTION

TABLE OF CONTENTS

The universe within which human behavior occurs is a variable one, with one important source of that variation being uncertainty or "error" in the data, that is, information received about some environmental source that does not reflect the true characteristics of that source. Therefore, situations in which the human perceiver must exercise judgment about the "true" value of a stimulus (whatever it's "apparent" value) are common.

Much of the error variation in the proximal representation of the environment available for processing can be conceived of as roughly Gaussian in character; the "true" value is obscured by a roughly normally distributed "error" component. In general, people are quite good at dealing with such variability. Brunswik (1956), for example, showed that in making perceptual judgments of size, people could successfully average out distance variations to attain "approximate size constancy." A similar conception underlies much of the research on probabilistic reasoning using numerical cues; here also "approximate cue constancy" can be attained if the errors are normally distributed around the true cue value. But what happens if the error is not Gaussian?

In a technology-driven ecology, much of the stimulation serving as the basis for judgments about distal objects or events is provided by fallible devices -- artifacts -- rather than by natural processes. In such situations true cues can be obscured by wildly non-Gaussian error components, as well as the more familiar Gaussian components. A calculator, for example, may give answers that are close approximations to true values most of the time, but give wildly wrong answers if the batteries are weak. How do individuals respond in such ecologies?

To examine this question, we have conducted a series of studies in

which error type is systematically manipulated. Before describing the research which has been accomplished under this contract, however, a brief discussion of how previous investigators have examined the effects of unreliability, or uncertainty in the data, is warranted.

The ability of people to deal with uncertainty has been of interest to experimental psychologists since choice reaction time was first systematically investigated. Formalizations of the concept of uncertainty in the 1950's in, for example, information theory, the theory of signal detectability and psychological decision theory, have made uncertainty one of the central concerns of cognitive psychologists.

While much research has been done on people's responses to environmental uncertainty, the great bulk of that research (outside the sensory domain) has dealt with uncertainty which has been pre-encoded by the experimenter. Hence, the subjects did not have to infer from observations of the data either the presence or degree of uncertainty. For example, in the many "bookbag-and-poker-chip" studies (Edwards, 1968) there was no uncertainty concerning what event had occurred; the subjects knew exactly what the sample proportion, say of red vs. white marbles, was. The uncertainty lay not in what the datum was but rather in the fact that the perfectly reliably observed datum provided probabilistic support for the hypotheses. The cascaded inference variant of the bookbag-and-poker-chip (Schum, 1977) work introduced such data uncertainty, but there too the uncertainty was pre-encoded by the experimenter and presented directly to the subject as observations which, while imperfectly diagnostic of some distal state, were themselves perfectly reliable.

Most of the research on the effects of "source reliability" encoded the uncertainty for subjects either by instructions, by semantic labels, or by presenting what subjects were expected to consider untrustworthy data (cf. York, Doherty and Kamouri, 1987, for a brief review). Most recently, in the "heuristics and biases" research, the environmental uncertainties have typically been pre-encoded as percentages, or described in verbal form (cf. Kahneman, Slovic & Tversky, 1982). In general, in the research alluded to above, perhaps most clearly in the work on source reliability, there is no way of knowing whether subjects are influenced by reliability or validity, in the technical psychometric senses of those terms.

In the present paper we will use reliability to refer to the degree to which a variable is stable over time, the degree to which a variable correlates with itself on repeated observations, or as the ratio of true score variance to total variance, where total variance is composed of true score and random error variance. Validity, of course, refers to the proportion of variance in common between a predictor and some criterion variable.

Relatively little is known about the effects of the form of uncertainty in which the data presented to subjects are, on a trial by trial basis, fallible, and such fallibility is either obvious to, or discoverable by, the subjects. Note again that by fallible we do not mean simply that the data are imperfect predictors of some other variable. We refer rather to the technical psychometric conception of unreliability in which a value of a variable is an imperfect indicator of the true state of that variable. We will use the term "data error" in lieu of unreliability, since we extend the definition somewhat beyond that typical in classical measurement theory.

THE NATURE OF DATA ERROR

In the classical psychometric conception of error, an observation, X_0, is composed of a true score component, X_t, and an error component, e.

$$X_0 = X_t + e \qquad (a)$$

The term e is usually considered to be a random variable sampled from a Gaussian distribution with mean zero. Its standard deviation, relative to that of X_t, determines the reliability of X_0, and influences the predictive and construct validity of the measure. This conception has proven to be a powerful approach to psychological measurement, and it is highly appropriate to a wide variety of measurement problems.

But classical error theory has been applied primarily to responses. The present research is concerned with the error in the data presented to subjects, the data on which subjects predicate their responses. There are forms of data error that are not typically differentiated by the classical psychometric conception, although they may have qualitatively different impacts on psychological processes. In particular, in many situations, data error may have an all-or-none quality; apparatus malfunction may produce wildly divergent meter readings on some apparently random proportion of trials, a physician may receive the results of a blood test on the wrong person, or a decimal error may be made. The distinction is more than that between categorical and continuous data, although errors associated with categorical data necessarily possess an all-or-none

quality. The distinction concerns rather the sort of error in which the datum at hand is simply unrelated to the causal or predictive process being studied. We will term error having this "all-or-none" quality "system failure" error (SF). The traditionally conceived Gaussian error we will call "measurement error" (ME).

As a framework for analysis, Figure 1 describes schematically a sequence of events typical of many studies of cognitive processes, and representative of at least some sequences of events in the world. The input data, X_{oi}, on which cognitive operations are performed, may not be perfectly valid or reliable. Hence, using the notation of (a) above,

$$X_{oi} = X_{ti} + e \qquad (1)$$

where i indexes the possible data inputs on any given event or trial. Example of ME applicable to organizations would include estimates on job applications of the durations of previous job tenures, rounded GPAs and ratings by the interviewer.

Analogously, the feedback provided to subjects may also be less than perfectly valid or reliable,

$$F_{oi} = F_{ti} + e \qquad (2)$$

where F_t is some true value. An employment interviewer may rarely get feedback about the performance of applicants he or she hired, but if such feedback were forthcoming, being told that such an employee was in

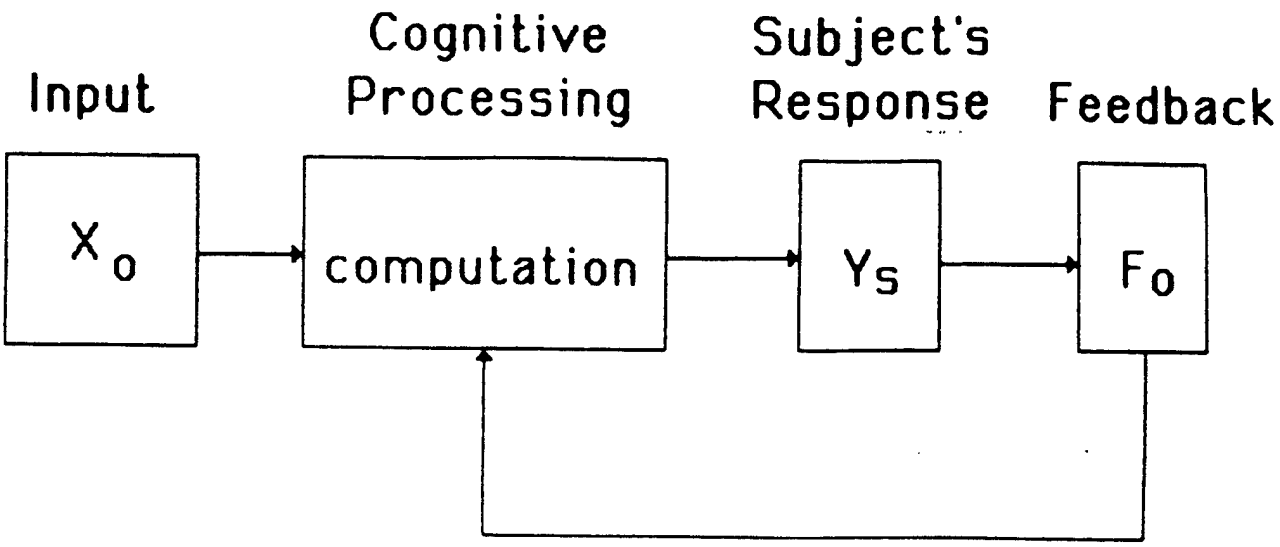


Figure 1. A typical trial in an experiment dealing with cognitive processes.

the top 10% when that salesman was actually in the top 15% would be in this category.

Of course, the determination of whether an error is a case of ME or SF is not possible given a single observation. Distributional information is essential to determine the nature of the error, in which the pattern of residuals from a correlational analyses in either reliability or validity analyses would be markedly different for data with SF than for data with ME errors.

In the case of SF, in the equation for data input error we introduce a different symbol for error, since it is a conceptually different sort of error. Let the observation that is unrelated to the true value of the variable putatively being measured be denoted E. Hence the equation for data input error with SF is

$$X_{oi} = E, \quad X_{oj} = X_t, \quad i \neq j \tag{3}$$

where values of E are selected randomly from the set of possible X_t, and substituted for the X_t that would have been presented on a specified trial. The feedback, F_o, is based on the weighted X_o called for in the experimental design on that trial. This equation shows, of course, that system failure is simply the limiting case of measurement error.

Similarly, the feedback error equation is, in the case of system failure,

$$F_{oi} = E, \quad F_{oj} = F_t, \quad i \neq j \tag{4}$$

where E is a random variable selected from the distribution of possible F_t without regard to the set of X_0 on that observation. Recalling the employment interviewer example, such a feedback error would occur if an interviewer had hired someone he or she thought would fail, then found out that the person had risen rapidly in the organization, never knowing that it was the nephew of the chairman of the board. A medical example would be a blood test on the wrong person when the physician is using the data to confirm an already made diagnosis. The possible psychological interest of all of the above enters, of course, when the person who is doing the diagnosis, employment interviewing, or whatever, becomes cognizant that the data being used are, in fact, subject to one or more of the error types described. The statistical relation between the proportion of SF errors and the population correlation coefficient is given in Doherty & Sullivan (In press, see Appendix 1).

Equations (3) and (4) may produce formally equivalent sets of relations between X_{0i} and F_0 under some circumstances. However, appropriate experimental manipulations may lead some subjects to regard the error as in the predictors, while others regard the error as in the feedback. Statistically equivalent amounts of error may have radically different psychological effects. Figure 2 summarizes the four error types.

A recent dissertation on the psychology of scientific inference (Kern, 1982) manipulated error type, and partly prompted the analysis shown in Fig. 2. Kern had advanced graduate students in the sciences make inferences based on data that were characterized by no error, ME error, SF error, or both forms of error. Her primary dependent variable was the

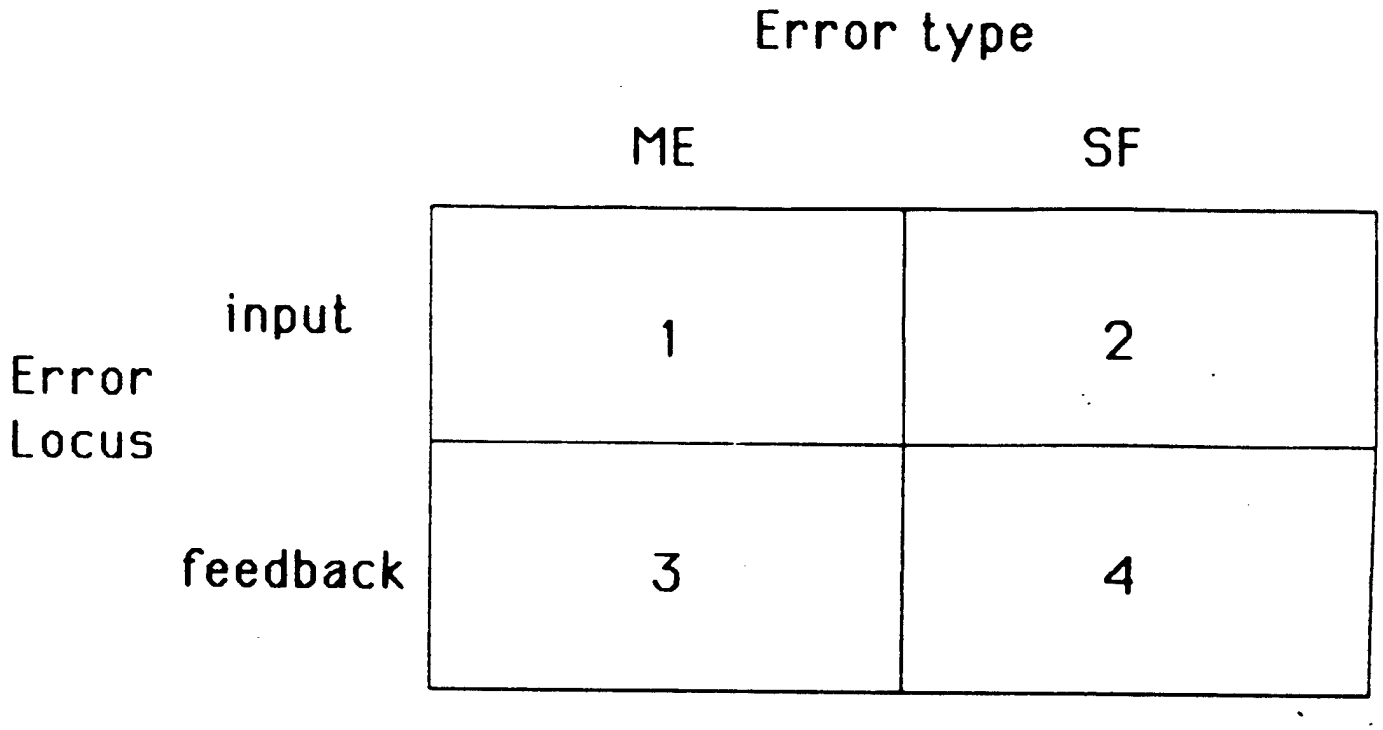


Figure 2. The four classes of error situations considered in this report.

tendency of subjects to change their hypotheses, given feedback. Subjects given measurement error changed hypotheses as readily as subjects with error-free data, while those subjects with system failure error displayed considerable hypothesis perseveration. In Kern's simulation of scientific hypothesis testing, the measurement error was essentially in the data on which the inferences were based, while the system failure error was in the feedback, i.e., she compared cells 1 and 4 in Fig. 2. The magnitude of effect in her work was truly impressive, suggesting the psychological importance of the distinctions among error categories, although not identifying which aspect(s) of the category system may be crucial. Data error has been manipulated by other investigators in various paradigms. Castellan (1977) describes a set of concept formation studies with "misinformative feedback". Any study with categorical input and/or feedback which gives subjects less than perfectly deterministic environments would qualify as representing SF error, given our definition.

The remaining investigations to be reported in Part 1 of the present paper have all been performed in the conceptual and analytic framework of the "lens model", which will be briefly presented.

THE LENS MODEL

The lens model has its roots in Probabilitic Functionalism, the system which represented Egon Brunswik's attempt to reshape both the theory and methodology of psychology (Hammond, 1966). Probabilistic Functionalism was cast primarily by Brunswik as a theory of perception (Postman and Tolman, 1959), but was subsequently broadened to the much wider domain of "human judgment" (Hammond, McClelland and Mumpower, 1980; Ullman and Doherty, 1984).

Brunswik argued that as far as the behaving organism could discover, the world presented a probabilistic, "semierratic" environment in which causes scattered their effects, an environment in which causes can be inferred only with some irreducible uncertainty from their effects. This conception is captured elegantly in a graphic representation, the lens model, shown in Figure 3. Suppose an investigator wishes to model some person's understanding of an environment. The essence of the method implied by the lens model, as extended especially by Hammond and his associates (we use essentially the notation of Hammond & Summers, 1972), is to have that person make a large number of quantitative or quantifiable judgments of multiattribute objects (scenarios, people, etc.), the attributes themselves being quantitative or quantifiable. The number of such objects which are judged must be sufficient to permit multiple regression analysis ("policy capturing") of each subject's data. It is this intensive statistical analysis of each subject's data that lends the seemingly paradoxical name "idiographic/statistical" to the approach.

Consider Figure 3. First note that the number of objects, or cases, to be judged by the individual is not represented. The attributes, hereafter to be referred to as cues, of each object are denoted by the X_{oi}. In the studies to be described in Part 1, there are two cues, but there is no reason to limit the number of cues to two: indeed as many as 64 have been used (Roose & Doherty.1976). For illustration, assume that a subject is judging 50 objects, each characterized by two cues. A possible task might be the selection of recruits for technical school based on two summary scores from a test battery. Such a task might be a pure "policy capturing" task, as when we assume that the person doing the evaluations already has some

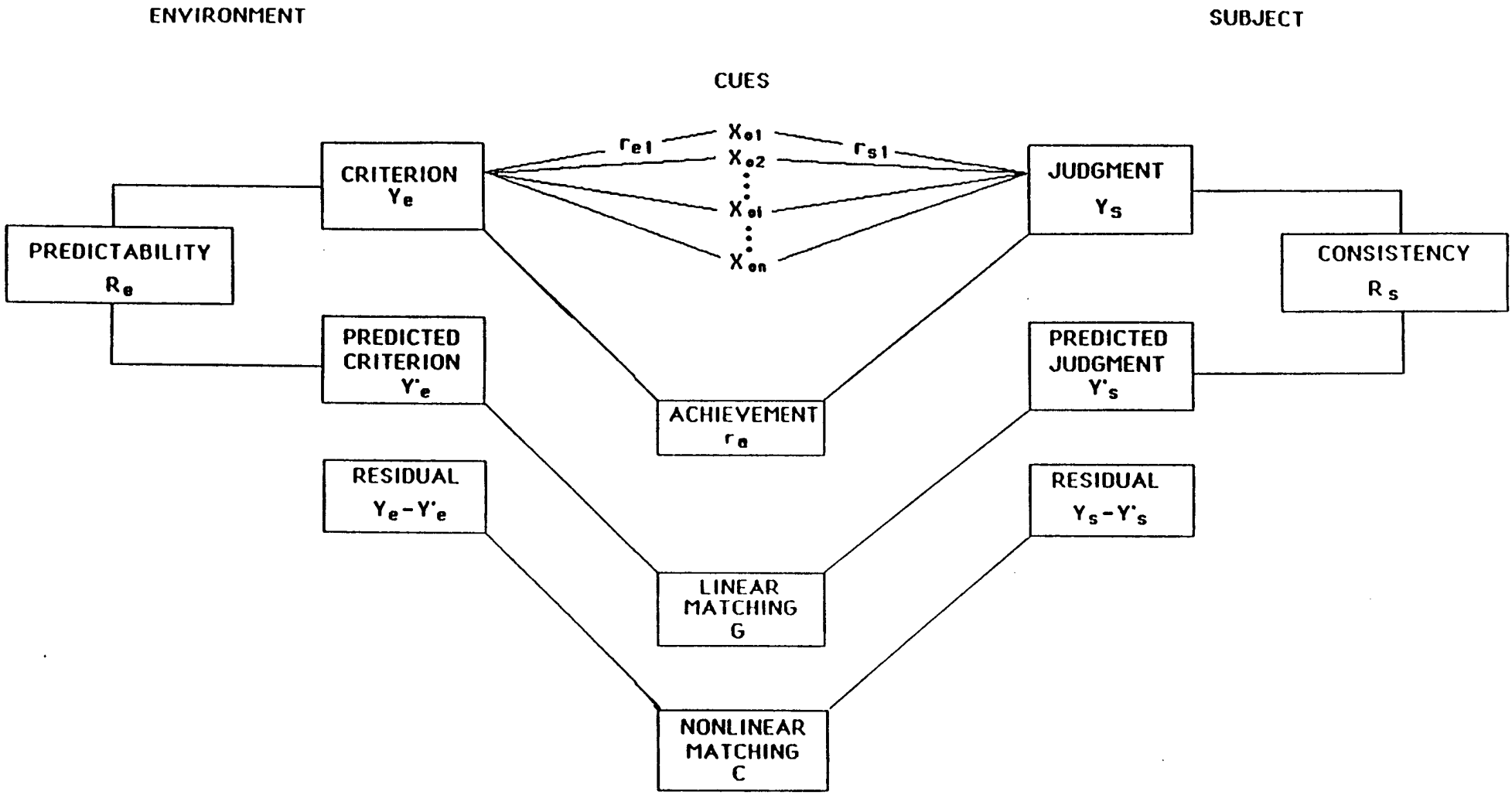


Figure 3. The lens model.

judgment policy which we are trying to discover and describe. In that case there would be no feedback to the subject, and no environmental side to the lens model. Or the task might be a learning task, in which the investigator is trying to discover how the subject comes to terms with the uncertainty inherent in a new prediction situation. Then both sides, or "systems" of the lens model, the environmental and the subject sides, would be needed to represent the situation. The Multiple Cue Probability Learning (MCPL) paradigm (Brehmer, 1980) exemplifies this latter application of the lens model. In the MCPL paradigm, the subject's task is to learn to predict a criterion variable from two or more cues. Generally a hypothetical situation is created. The cues are presented, the subject predicts a criterion value and outcome feedback is presented. This procedure, which follows that outlined in Figure 1, may be followed for 100 or more trials, with the cues being only probabilistically related to the criterion, and the criterion typically not being perfectly predictable even with optimal weighting and combining of the cues. People do not do well in such a task, without help.

The lens model depicts the simultaneous relationships of the cues with both the "ecological", or criterion, variable and with the judgment variable. For example the cue X_{01} possesses a relationship to the criterion, Y_e, described by the correlation coefficient, r_{e1}, which correlation is called the ecological validity of that cue. Similarly X_{01} possesses a relationship to the judgment, Y_s, described by the correlation coefficient, r_{s1}, called the utilization coefficient. Optimally the values of the utilization coefficients would match the values of their respective

ecological validities. In the world, this simply does not occur, and it is in the attempt to understand the probabilistic behavior of people in a probabilistic world that the lens model and the associated " lens model equation" are useful tools. Consider the variation of the lens model equation developed by Tucker (1964) from the work of Hammond, Hursch & Todd (1964) and Hursch, Hammond & Hursch (1964):

$$r_a = R_e R_s G + C[(1-R_e^2)(1-R_s^2)]^{1/2}$$

The most straightforward way to understand the components of the lens model equation are as bivariate correlations:

r_a the correlation between the subject's judgments (Y_s) and the actual values of the criterion (Y_e): the "achievement" index.

R_e the correlation between the actual values of the criterion (Y_e) and the values predicted by the multiple regression analysis (Y'_e): alternatively the multiple R between the cues and the criterion variable. This reflects how predictable the environment is, given the assumption of a linear, additive model.

R_s the correlation between the actual values of the judgments (Y_s) and the values predicted by the multiple regression analysis (Y'_s), alternatively the multiple R between the cues and the judgment variable. This reflects how predictable, or "consistent" the subject is, given the assumption of a linear, additive model.

G the correlation between the values of the criterion variable predicted by the multiple regression analysis on the environmental side (Y'_e) and the values of the judgment variable predicted by the regression analysis on the subject side (Y'_s). This reflects the knowledge the subject has about the linear, additive properties of the environment.

C the correlation between the residuals on the environmental side of the lens (Y_e-Y'_e) and the residuals on the subject side of the lens (Y_S-Y'_S). This reflects, among other things, the subject's knowledge of the nonlinearities and nonadditivities in the environment.

The lens model equation is a rigorous mathematical expression of an elementary truism: the ability of a person to predict the world depends on how predictable the world is, how consistently the person processes information about the world, and how well the person understands the world. If any one of these components is lacking, prediction is impossible: predictability in the equation is fundamentally a product of three decimals, and if any one is zero the product is zero (empirically, the additive component is typically vanishingly small).

Note that in Figure 3 the notation representing the criterion variable is Y_e. This corresponds to F_t in Figure 1, the different notation for the same value serving to highlight that in a given experiment the Y_e is not only the value of the criterion variable but is also being used in the experimental setting as the basis of the feedback.

The present research uses extensions of the MCPL paradigm, not to study learning per se, but as a relatively well-understood vehicle to study possible effects of data error. It is clear from existing research that subjects in a typical MCPL study can learn simple cue-criterion relations from repeated pairings of multiple cues with the criterion given as feedback. It is also clear that such learning is inefficient compared with more cognitively oriented feedback, especially in more complex environments (Hammond, 1986; Hammond, McClelland & Mumpower, 1980).

PART 1

RESEARCH ON DATA ERROR IN THE LENS MODEL PARADIGM

CONTENTS

A. PRIOR RESEARCH

Brehmer, 1970

Brehmer (1970) manipulated reliability in a series of studies by having subjects infer their own values of the predictor variables. Subjects attempted to predict the meeting place of two automobiles from their perceptions of the velocities of the two vehicles, one in which they rode and the other which came from the opposite direction. While the task was perfectly predictable given the actual velocities, this self-generated unreliability was such as to reduce the task predictability to less than a multiple correlation of 1.0. The essential conclusion of this research was that people treat self-generated uncertainty about the cue values much as they treat the uncertainty inherent in less than perfectly predictable relationships between cues and criteria in standard MCPL studies. One of the major sources of evidence for this generalization was the tendency of subjects to match R_s to R_e. Note that the use of self-generated unreliability still confounds reliability and validity.

Markowitz, 1983

Since this investigation, which is relevant to the purposes of the contract, is unpublished, we will present it in some detail. Markowitz' (1983) study was a variation of Brehmer's (1971) investigation of self-generated unreliability. Subjects used objective measurements and/or perceptual estimates of the heights of two "sisters" to predict the

height of a third, unseen sister. Since people might not spontaneously distinguish between reliable measurements made with a scale and relatively unreliable perceptual estimates, one of the independent variables was an instructional manipulation of the salience of the reliability differentials. A second variable was whether the perceptual estimates were or were not recorded, and a third was the number of unreliable predictors, one or two.

Color slides of 80 females were taken from a uniform distance of about 3.5 m., against a plain background. The only cue to height was a sidechair of standard height, which was placed in front of each female such that one of her legs was visible. Forty-eight slides were selected from the 80 and divided into two halves so that the persons in each pair appeared to be sisters, and the heights in each group were uncorrelated, normally distributed, and had equal means and variances. The actual value of the r between the true heights was -.02, and the standard deviations of the true heights were 2.34 and 2.56 inches for the first and second sisters, respectively.

Subjects participated in small groups in one of nine conditions. They were instructed that they would see 44 pairs of slides, each slide with a photograph of one sister, and that their task would be to estimate the height of a third sister from the heights of the first two. The nine conditions were produced by a 2 X 2 X 2 factorial combination of three independent variables plus a Control (cf. Himmelfarb, 1975). The Control group was given the heights of both sisters, as "obtained by the experimenter, using a scale". Half the remaining 80 subjects had the heights of the second sisters already on the response sheet: the other half

had to estimate both heights (one reliable cue vs. two unreliable cues-RU vs. UU, with R and U denoting reliable and unreliable.). Half the subjects were instructed that they "should have less confidence in heights that you estimate than in heights that have been measured on a scale": the other half had no instructions highlighting the unreliability of their estimates (instructions vs. not instructed-I vs. NI). Half the subjects wrote down their estimates: half did not (written vs. not written estimates-W vs. NW).

Subjects saw four practice pairs, 20 actual pairs, took a one min break, then saw the 20 actual pairs in a different random order. For each pair they predicted the height of the third sister, and made the estimate(s) called for by the condition to which they had been assigned. The heights were written by subjects on a line or lines provided on that same sheet with the measured heights appropriate to their experimental condition.

Normatively, estimated heights should be less variable than actual heights (i.e., the estimates should be regressed to the mean) as should predictions of the third sister's height. The predictions should be regressed least given two reliable heights, and most given two estimated heights. Further, the weights ascribed to the reliable heights ought to be greater than those to the unreliable heights. Also, the multiple R relating the predicted height to the cues should not vary with the reliability of the cues, given purely normative considerations.

The data of this study are summarized briefly in Table 1-1.

Table 1-1

Means from the regression analyses in Markowitz' dissertation.[1]

	Statistical indices of performance						
Condition	SD'1	SD'2	r'_{12}	PC'2-PC'1	R'_S	PC2-PC1	R_S
CONTROL	n/a	n/a	n/a	n/a	n/a	.40	.85
RU,NI,W	2.00	n/a	.52	.22	.94	.78	.89
RU,NI,NW	n/a	n/a	n/a	n/a	n/a	.86	.91
RU,I,W	1.87	n/a	.48	.22	.91	.69	.88
RU,I,NW	n/a	n/a	n/a	n/a	n/a	.85	.93
UU,NI,W	2.25	2.49	.60	.38	.94	.16	.72
UU,NI,NW	n/a	n/a	n/a	n/a	n/a	.20	.69
UU,I,W	2.21	2.26	.53	.27	.89	.21	.60
UU,I,NW	n/a	n/a	n/a	n/a	n/a	.19	.69

[1]The abbreviations for the conditions are defined in the text. The primes on the statistical indices indicate that the analyses were performed on predictors with either one or two subjective estimates of height, depending on the condition. SD refers to the standard deviation of the estimates, r'_{12} to the correlation between predictors, and PC2-PC1 to the difference between partial correlations between predictors and predictions. The last two columns provide the mean difference between partial correlations and the multiple R given the true heights as predictors. The symbol n/a indicates that a value would be inappropriate in that cell in the table. For example, there is no value for SD'1 since the SD of the first sister's height was fixed by the experimenter, and is not a performance index.

Detailed statistical analyses can be found in Markowitz (1983). The overriding implication of the above results and of the written reports made by the subjects is that while the unreliability of the input data upon which predictions are based does influence aspects of prediction behavior, those influences are not systematically in line with what one would expect from normative theory. Subjects are not appropriately sensitive to error in input data, not even remotely so. Although a degree of regressiveness was noted, as would be partly expected from other research demonstrating insufficient regressiveness (Kahneman, Slovic & Tversky, 1983), the regressiveness of estimates and predictions was neither systematically present nor necessarily appropriate when present.

There is a tangential issue of some practical interest: writing down the estimates led to less normative behavior. It seems that once the estimate was written down it was used no differently than a value obtained from a more reliable source. The implication of this finding appears to run counter to much conventional wisdom, which exhorts raters, decision makers, etc. to record their observations lest forgetting should cause a loss of the information. But if observations which are error prone (i.e., all observations) take on the aura of a precise number when recorded, then the failure to account for unreliability may have consequences worse than the loss of the observation due to forgetting would have had, should the recorded observation later be aggregated with some more reliable data. Perhaps a fruitful avenue of research to pursue would be the effects of various methods of having the observer record some estimate of the quality of the datum at the same time as the datum itself is recorded.

The present study did not replicate in detail the finding of Brehmer (1970) that the consistency of predictions was reduced in response to self-generated unreliability in the predictors. The experiments differed in so many particulars that there are no reasonable grounds for speculating about this difference. Nor do we have any grounds for speculating about the utter lack of effect of the instructions to attend to the unreliability of the estimates in either the statistical analyses or the subjective reports.

There are two features of the study which detract from otherwise quite comprehensible results. One is the construction of the data set in such a fashion that the correlation between the true heights of the "sisters" was almost exactly zero. This was done for pragmatic reasons, i.e., the difficulty of getting a sufficient number of true sisters, and the desire to maximize the interpretability of dissertation results. This procedure runs counter, of course, to Brunswik's (1956) call for the representative design of experiments. The other difficulty, which this task shares with Brehmer's (1970) work, is that the "sisters" task does not neatly separate the reliability from the validity of the input data.

The remainder of this part reports a series of studies, each bearing on some aspect of data error, each using some variant of the MCPL approach. The particulars of the studies differ a great deal, partly because different investigators had different conceptions of how error might best be manipulated, but largely because of a conviction that a variety of tasks should be used in order to make it possible to generalize across tasks.

B. RESEARCH CONDUCTED UNDER THIS CONTRACT

EXPERIMENT 1

York, Doherty & Kamouri (1987) assessed the impact of ME in the input. Error was manipulated by having two computer-simulated "meters" provide multiple observations of two underlying variables. The meters varied in the degree to which they "jittered". The presence of such measurement error did not have interesting effects on subjects' cue utilization, consistency or achievement. York et al. concluded that the subjects were essentially averaging over the measurement error, and hence were engaging in an appropriate strategy to deal with the complications introduced by the random variation of the observed scores. The highly salient presence of measurement error did, however, influence subjects' self-reports about the degree to which they attended to the cues. This study has been published, and a reprint is provided in the appendix.

EXPERIMENT 2

This experiment was also designed to investigate the impact on predictions of making it transparent to subjects that input data are not reliable. A standard MCPL condition, with an uncertain environment but with the source of that uncertainty unspecified, served as a baseline condition. There were two experimental conditions, one with the true values of the cues provided after the subjects had made their predictions, the other with multiple observations of the cues used for prediction. If the subjects are not influenced by ME in the input, then the former experimental condition should be no better or worse than controls, and

the latter, multiple observation condition, should be much better. Conversely, if the error leads people to distrust data, then the performance of the experimental subjects may be worse than that of the controls.

METHOD

Subjects. A total of 46 jet fighter pilots rated to fly F-15 and F-16 aircraft, stationed at a USAF base in the continental United States, participated as volunteers in this experiment. They received no tangible compensation of any sort. Most of the participants were captains, with a few lieutenants and majors.

Apparatus. All stimulus materials, including instructions and pages for 104 MCPL trials, two pages per trial, were in 3 inch loose leaf binders. The forms for the trials were Xeroxed on 8 1/2 in, 67 lb white Vellum Bristol-Cover stock, and the data specific to a trial were entered by hand with a ball point pen. These materials are described below, and shown in Appendix 1. There were two experimental manipulations and a standard MCPL treatment. The latter, which will be referred to as the CONTROL treatment, will be described first.

CONTROL. The instructions informed the pilots that the experiment was a study in "how people use data", that they were to play the role of an executive trainee whose task was to predict the weekly amount of energy consumed by an organization. Each prediction was to be based on two forecasts, a forecast of temperature and a forecast of the number of units produced. They were also told that temperature and units produced were unrelated, and told in simple terms that both were related positively to energy consumption.

In the binder, each trial was represented by two consecutive pages. At the top of the first page for each trial was the label " Data for Forecast Week No. ___" Below that was a figure which filled roughly the middle half of the page. The figure had a horizontal baseline approximately 18 cm across the bottom, and two vertical columns, 9.5 cm high and 1 cm wide, the left sides of which were approximately 1 and 4 cm from the left edge of the baseline, leaving 2 cm between columns. Across the width of the column was a horizontal line, the height of which represented the cue value. One bar was labeled "Temp.", the other "Prod", for temperature and units produced, respectively. At the bottom of the page in large type were the lines

Temperature Forecast _____ °F

Production Forecast _____ units

Entered on these blank lines were the values of the temperature and production forecasts. These values were perfectly redundant with the heights of the horizontal lines drawn in the columns. Subjects inspected the values, made their predictions on a response sheet, then turned the page and read the "End of Week Report Week No. __ " This page was identical to the first except that there was added a third column, labeled "Energy" and a third line at the bottom of the page, "Actual Energy Consumption _____ units." As with the cues, a line in the Energy column and a redundant number in the bottom line represented the criterion value, i.e., F_0 in the notation developed in the introduction. Even though heavy card stock had been used, the information on page 2 would have been detectable when looking at page 1, i.e., through the input page. Therefore, before subjects were run, the backs of all copies of page 1 had

large black rectangles in appropriate places, rendering those areas opaque.

This procedure was followed for 79 trials, the first four of which were practice trials which were dropped from the analysis. The last 25 trials were preceded by a page in the binder which read: "The next set of 25 trials is a test set to measure what you have already learned. There are no more end of week reports". Such a test block without outcome feedback is typical of many MCPL studies. It is an effort to obtain performance indices uncontaminated by policy shifts made as a consequence of feedback. The entire sequence of trials took about an hour to complete, after which a brief post-experimental questionnaire was administered.

While the order of cues on the pages was constant, the cue validities were randomly assigned. Thus, "Temp" was always on the left, but for any given subject temperature might have had either the higher or the lower correlation with the criterion.

The statistical structure of the task was controlled by creating arrays of random numbers in a computer, factor analyzing the random arrays, then performing the appropriate transformations on the resulting zero-correlated factor scores to achieve the desired formal task characteristics. After rounding to two digits, the two cues correlated .002. The ecological validity (i.e., the correlation between the cue and the criterion) of X_{01} was .54 and of X_{02} was .58, with the subscripts 1 and 2 not denoting order but simply serving as identifiers for the sake of presentation in this report. Task predictability (R_e) was .63, both over the entire set of 104 trials and the set of 79 that included Y_e. The statistical structure of the task is summarized in Table 1-2. Pages 29-36 provide

the complete instructions, a sample "week", the data sheet and the brief post-experimental questionnaire for this condition.

TABLE 1-2

The statistical structure of the task environment in Experiment 2.[1]

	X_{t1}	X_{t2}	X_{o1a}	X_{o1b}	X_{o2a}	X_{o2b}	Y_e
X_{t1}	100	00	95	95	00	00	57
X_{t2}		100	00	00	71	71	82
X_{o1a}			100	90	00	00	54
X_{o1b}				100	00	00	55
X_{o2a}					100	50	58
X_{o2b}						100	58
Y_e							100

[1]Decimals omitted. X denotes the cues and Y_e the criterion, the subscripts t and o denote true and observed, the subscripts 1 and 2 represent the particular combination of validity and reliability, and the subscripts a and b the separate observations of a cue in the multiple observation treatment.

INSTRUCTIONS

In this research we are interested in how people use data. To investigate data usage, we are asking a large number of people to participate in a variety of business simulations. The data for your business simulation is in the large binder. The specific instructions are on the following pages.

Figure 1-1. Pages 29 -36 provide sample pages from the CONTROL condition of experiment 1.

SIMULATION

Prediction of Weekly Energy Consumption

Superior Industries, Inc.

You have just begun work as an executive trainee for a manufacturing corporation. One of your jobs is to report to the financial planning executive concerned with short-range financial issues.

One of the major costs in running a company is energy consumption and each week you will be required to predict how much energy the company will consume. To get you acquainted with making these predictions you are given weekly records for the past two years and asked to make projections of energy consumption for each of those 104 weeks.

You will be given information concerning the two factors that determine energy consumption. On the basis of these two factors you will make a prediction of the weekly energy consumption.

The two factors that determine energy consumption are:

(1) Temperature, and

(2) Number of units manufactured.

However, nobody knows exactly what the weather will be or exactly how many units will be produced in the coming week. This means you will have to make your predictions based on:

(1) Weather forecasts, and

(2) Forecasts of units to be manufactured.

The weather forecast for each week will range from about 20 degrees to about 75 degrees. The forecast of units to be manufactured for each week, based on sales projections, shipment schedules, and available labor, will vary from about 20 to 75.

The data for the two weekly forecasts, weather and units to be manufactured, will be on a single page in the binder. For your convenience they will be given as numbers and as bar graphs. You are to predict energy consumption based on these two forecasts. The range of energy consumption is from about 350 to 850 units.

After making your prediction of energy consumption on the data sheet provided, turn the page where you will find the actual energy consumption for that week.

Your first predictions probably will be guesses. However, as you work through the weeks you will be able to learn the relationships between the two forecasts and energy consumption.

This task has been arranged so that you can learn to predict energy consumption with moderate accuracy. As in the real world, though, predictions can never be perfectly accurate all the time.

Your task is to make the closest predictions possible. This is a difficult task, so we would like to tell you some things about it. There is no relationship at all between temperature and number of units manufactured, but both influence energy consumption in a direct fashion. That is, as temperature goes up energy goes up and as units manufactured goes up energy goes up. But temperature and units manufactured are unrelated.

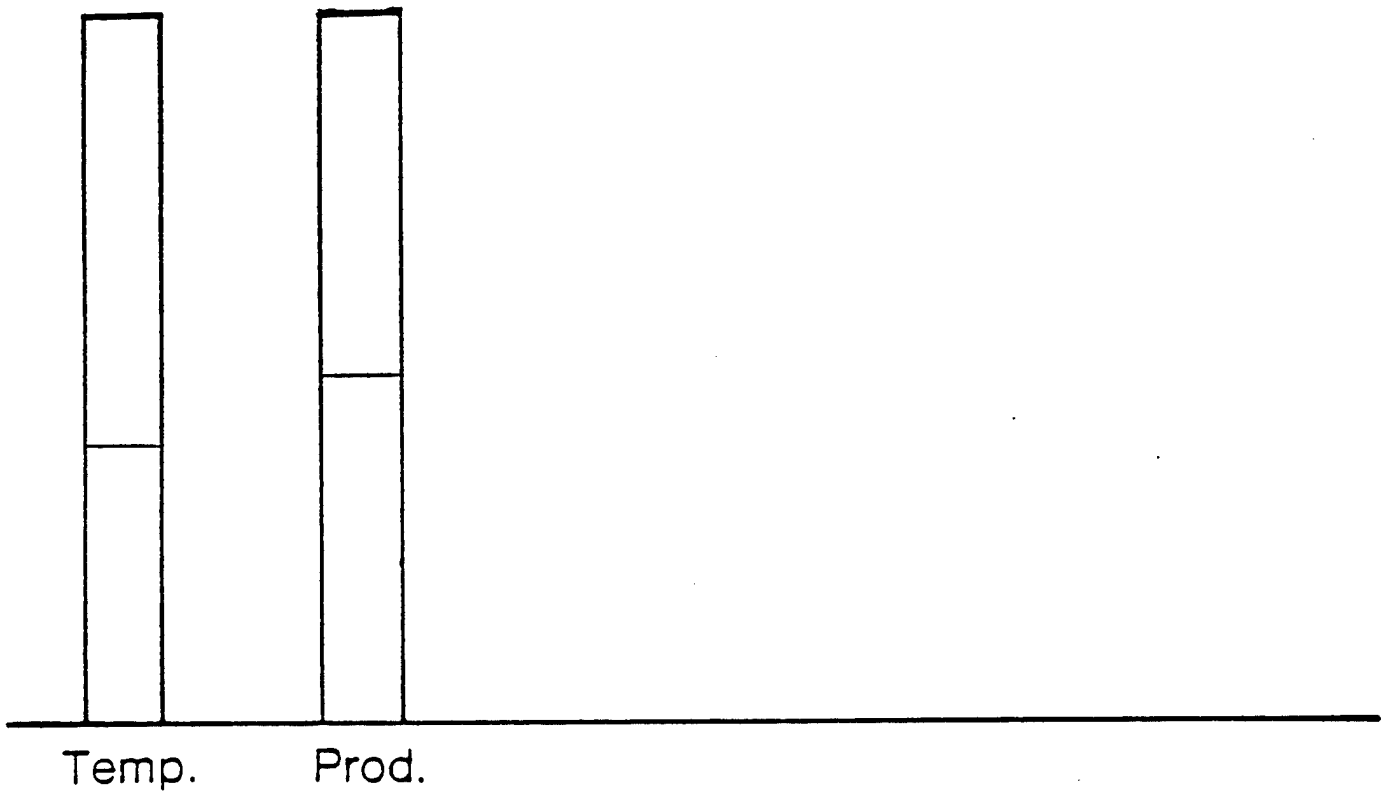
The first 4 weeks are practice. Do those first and if you have any questions ask them after you've completed these first 4. Weeks 5 - 79 are the learning trials. On each of these you make your prediction, then turn the page and find out actual energy consumption.

Weeks 80 - 104 are test weeks. On these you will not get feedback about the correct values. After the 104th week, you will be asked whether the temperature forecast or the production forecast provides more useful information for prediction of energy consumption.

Please go ahead.

Data for Forecast

Week No. 31

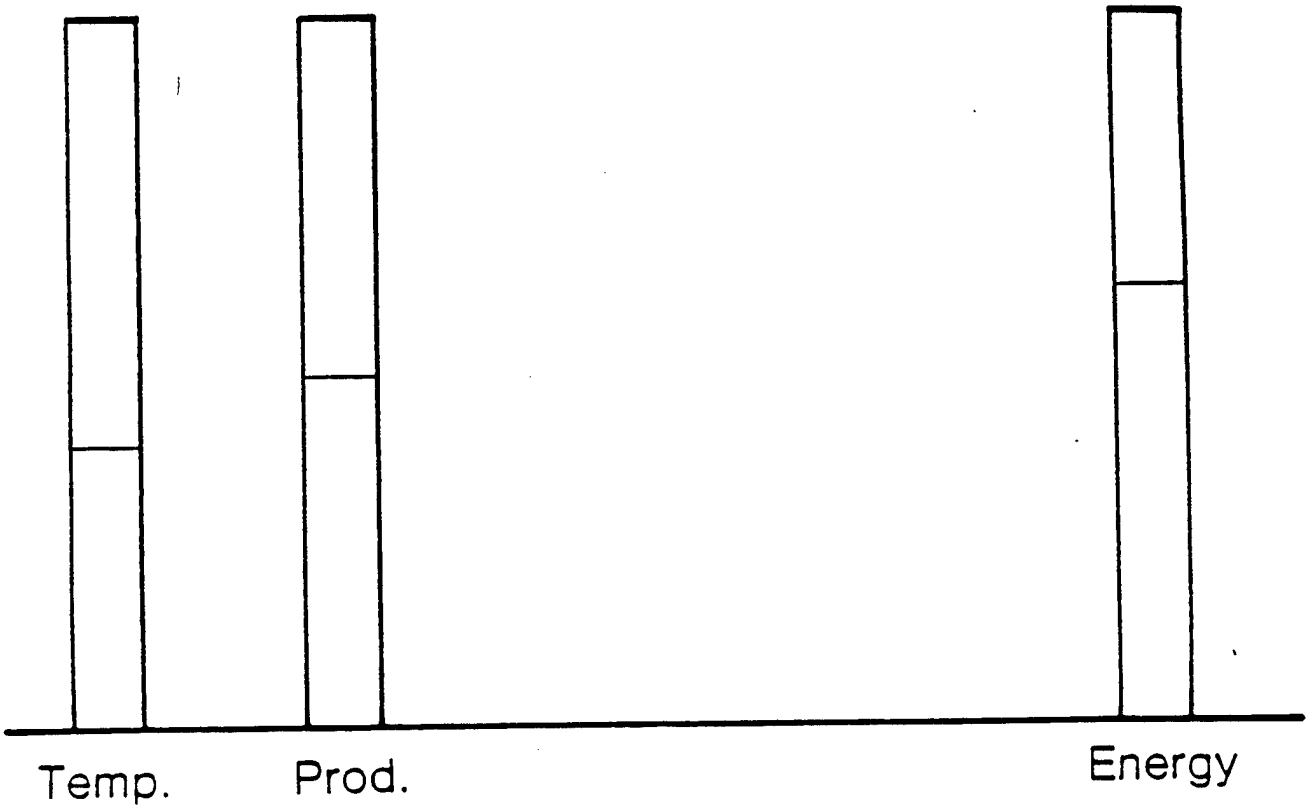


Temperature Forecast 36 °F

Production Forecast 45 units

End of Week Report

Week No. 31



Temperature Forecast 36 °F

Production Forecast 45 units

Actual Energy Consumption 362 units

Prediction of Weekly Energy Consumption

Superior Industries, Inc.

DATA SHEET

Practice Weeks	Training Weeks	Training Weeks	Test Weeks
1. _____	30. _____	55. _____	80. _____
2. _____	31. _____	56. _____	81. _____
3. _____	32. _____	57. _____	82. _____
4. _____	33. _____	58. _____	83. _____
Training Weeks	34. _____	59. _____	84. _____
5. _____	35. _____	60. _____	85. _____
6. _____	36. _____	61. _____	86. _____
7. _____	37. _____	62. _____	87. _____
8. _____	38. _____	63. _____	88. _____
9. _____	39. _____	64. _____	89. _____
10. _____	40. _____	65. _____	90. _____
11. _____	41. _____	66. _____	91. _____
12. _____	42. _____	67. _____	92. _____
13. _____	43. _____	68. _____	93. _____
14. _____	44. _____	69. _____	94. _____
15. _____	45. _____	70. _____	95. _____
16. _____	46. _____	71. _____	96. _____
17. _____	47. _____	72. _____	97. _____
18. _____	48. _____	73. _____	98. _____
19. _____	49. _____	74. _____	99. _____
20. _____	50. _____	75. _____	100. _____
31. _____	51. _____	76. _____	101. _____
22. _____	52. _____	77. _____	102. _____
23. _____	53. _____	78. _____	103. _____
24. _____	54. _____	79. _____	104. _____
25. _____			
26. _____			
27. _____			
28. _____			
29. _____			

Post-Experimental Questionnaire

1. If you had to predict energy consumption based on either the weather forecast or the production forecast, which would you choose?

Check one: Weather Forecasts _____ Production Forecasts _____

2. Please try to describe the thinking process you went through in making your predictions. How did you go about trying to learn and use the relationships between temperature and production and the level of energy consumption?

INPUT ERROR SHOWN. In the above condition, which was a standard MCPL paradigm, there was no way for the pilots to determine the locus of the error, or even if there were any error, since the lack of perfect predictability of the environment could have been due to an unmeasured variable or variables. The experimental manipulation in the INPUT ERROR SHOWN treatment was designed to make it self-evident that the error in the system was in the forecasts, not in the criterion values. This was accomplished by showing the true values of the cues on every trial, along with a criterion value which was perfectly predictable from those values, after the forecast of energy consumption had been made by the subject.

The materials were identical with the CONTROL materials except as noted. After making the forecast of energy consumption, the subject turned to the feedback page and got, in addition to the information provided to the CONTROL subjects, the values of X_t. These were presented as red lines in the cue bars, at the appropriate heights above the baseline, and as red numbers written at the bottom of the page on lines labeled "Actual Temperature" and "Actual Production." These entries on the feedback page were immediately below the respective repetitions of the values of the forecasts. To be consistent for the subjects, the line and numerical entry for the "Actual Energy Consumption" were also in red. The final difference, other than the implied changes in the instructions, was that the top of each feedback page had a prominent notation explaining the color coding.

The statistical structure of the INPUT ERROR SHOWN condition was the same as for the CONTROL condition, with the following additional

information concerning the true scores. The r between the X_{ti} was -.002, between X_{t1} and Y_e was the r was .57, and between X_{t2} and Y_e the r was .82. The correlation between X_{t1} and X_{01} (i.e., the reliability of X_{01}) was .95, and that between between X_{t2} and X_{02} was .71. Hence in this condition, the more reliable cue was the less valid cue, and vice versa. Note also that in this condition the criterion would be perfectly predictable, were one given the true values of the predictors.

Pages 39-41 provides an example of an "End of Week Report" and two pages of the instructions. The other pages, including the "Data for Forecast" pages, were identical to those in the CONTROL condition.

After making your prediction of energy consumption on the data sheet provided, turn the page where you will find:

(1) The actual average temperature for that week,

(2) The actual number of units manufactured for that week, and

(3) The actual energy consumption for that week.

Notice that the forecast of temperature may have been different from the actual temperature and that the forecast of units may have been different from the actual units.

Your first predictions probably will be guesses. However, as you work through the weeks you will be able to learn the relationships between the two forecasts and energy consumption. You will also learn how accurate the weather forecast is and how accurate the manufacturing forecast is.

This task has been arranged so that you can learn to predict energy consumption with moderate accuracy. As in the real world, though, predictions can never be perfectly accurate all the time.

Your task is to make the closest predictions possible. This is a difficult task, so we would like to tell you some things about it. There is no relationship at all between temperature and number of units manufactured, but both influence energy consumption in a direct fashion. That is, as temperature goes up energy goes up and as units manufactured goes up energy goes up. But temperature and units manufactured are unrelated.

The first 4 weeks are practice. Do those first and if you have any questions ask them after you've completed these first 4. Weeks 5 - 79 are the learning trials. On each of these you make your prediction, then turn the page and find out actual temperature and production values, as well as the actual energy consumption.

Weeks 80 - 104 are test weeks. On these you will not get feedback about the correct values. After the 104th week, you will be asked whether the temperature forecast or the production forecast provides more useful information

Figure 1-2. Pages from the INPUT ERROR SHOWN condition of experiment 1.

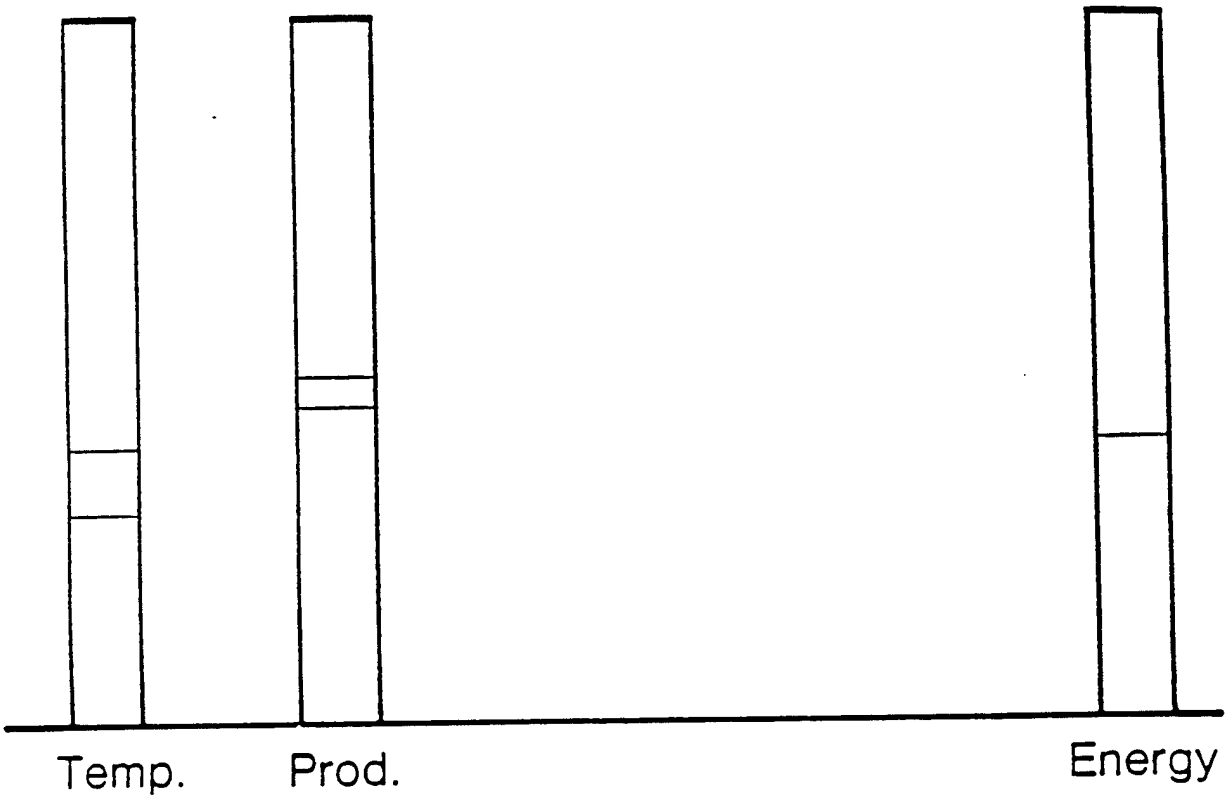
for prediction of energy consumption. You will also be asked how accurate the forecasts themselves are; that is, how well the temperature forecasts predicted actual temperature versus how well the production forecasts predicted actual production.

Please go ahead.

End of Week Report

Week No. 31

Actual Data Shown In Red
Forecast Data Shown In Blue



Temperature Forecast 36 °F

Actual Temperature 27 °F

Production Forecast 45 units

Actual Production 41 units

Actual Energy Consumption 362 units

MULTIPLE OBSERVATIONS. In this condition, the subjects were also "hit between the eyes" with the unreliability of the cues on every trial. This was accomplished by presenting subjects with two different forecasts on each trial for both weather and production. The pilots were informed that in this simulation they had weather forecasts for each week from both the "National Weather Service" and a "Staff Meteorologist", and Production forecasts from both the "Production Manager" and the "Quality Control Supervisor" (see pp. 43-44). The binder pages (see pp. 45-46) now had four cue columns where before there had been two, and four numerical values entered at the bottom of the page, with suitable modifications of the page. Thus, whereas the pilots in the the INPUT ERROR SHOWN condition were shown error after they made each prediction, in this condition the information about the locus of the error was given before they made each of their predictions. The correlational structure of the task was as in the INPUT ERROR SHOWN condition, with the additional feature that the two sets of X_{0i} were themselves correlated. As can be inferred by squaring the appropriate correlations between true and observed cues described above, the correlation between the two values of X_{01} was .90, and .50 between the two values of X_{02}. Note that giving the subjects two independent observations of each cue necessarily involves confounding the multiple observation manipulation with either cue validity or task predictability. The latter confounding was present in this task, with R_e for this condition being .76, which was of course substantially higher than in the other two conditions. Hence, subjects in this condition had a higher theoretical performance ceiling than in either of the other two conditions.

Note that there are two separate forecasts for temperature, and two separate forecasts for production. The separate forecasts are from different sources and agree with one another reasonably well, but not perfectly. Two of the sources agree more closely with each other than do the other two.

After making your prediction of energy consumption on the data sheet provided, turn the page where you will find the actual energy consumption for that week.

Your first predictions probably will be guesses. However, as you work through the weeks you will be able to learn the relationships between the two forecasts and energy consumption.

This task has been arranged so that you can learn to predict energy consumption with moderate accuracy. As in the real world, though, predictions can never be perfectly accurate all the time.

Your task is to make the closest predictions possible. This is a difficult task, so we would like to tell you some things about it. There is no relationship at all between temperature and number of units manufactured, but both influence energy consumption in a direct fashion. That is, as temperature goes up energy goes up and as units manufactured goes up energy goes up. But temperature and units manufactured are unrelated.

The first 4 weeks are practice. Do those first and if you have any questions ask them after you've completed these first 4. Weeks 5 - 79 are the learning trials. On each of these you make your prediction, then turn the page and find out actual energy consumption. For your convenience, the forecast data are repeated on the "end of week report."

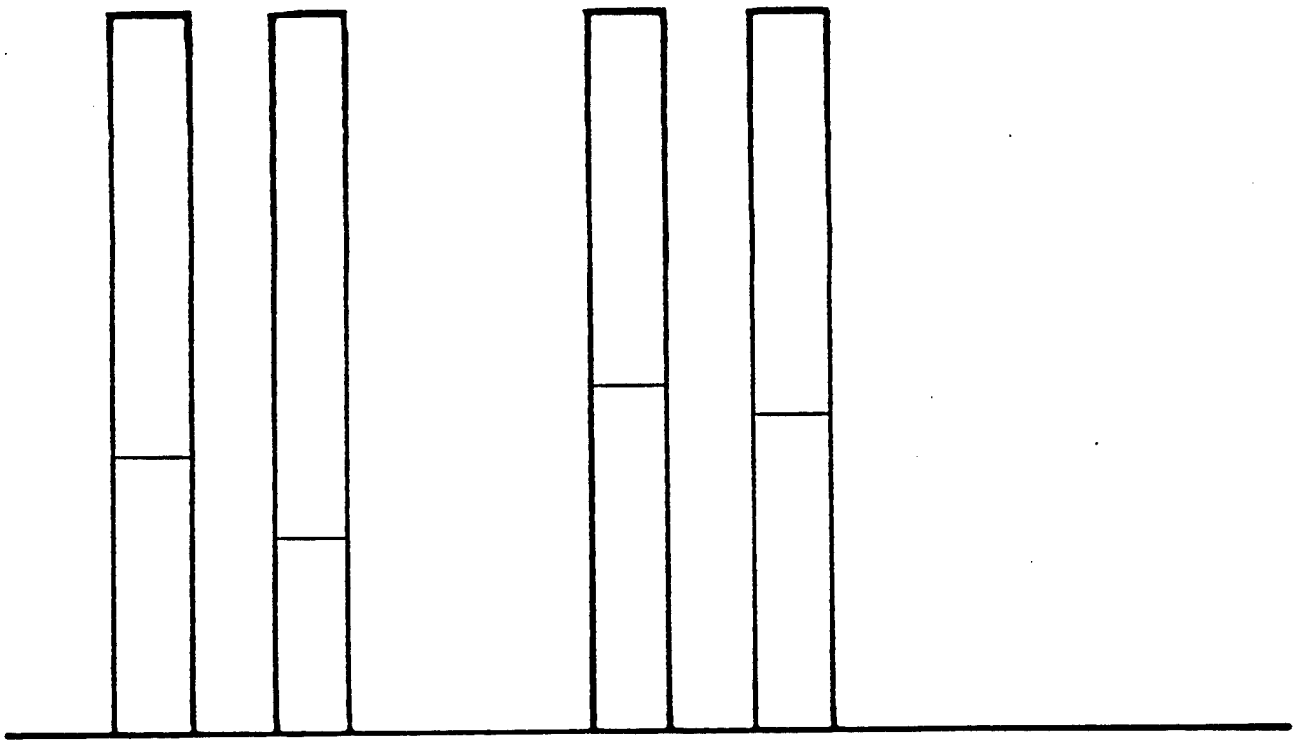
Figure 1-3. Pages from the MULTIPLE OBSERVATION condition.

Weeks 80 - 104 are test weeks. On these you will not get feedback about the correct value of energy consumption. After the 104th week, you will be asked whether the temperature forecast or the production forecast provides more useful information for prediction of energy consumption. You will also be asked which sources of the forecasts agreed better between themselves, the two sources for temperature or the two sources for production.

Please go ahead.

Data for Forecast

Week No. 31

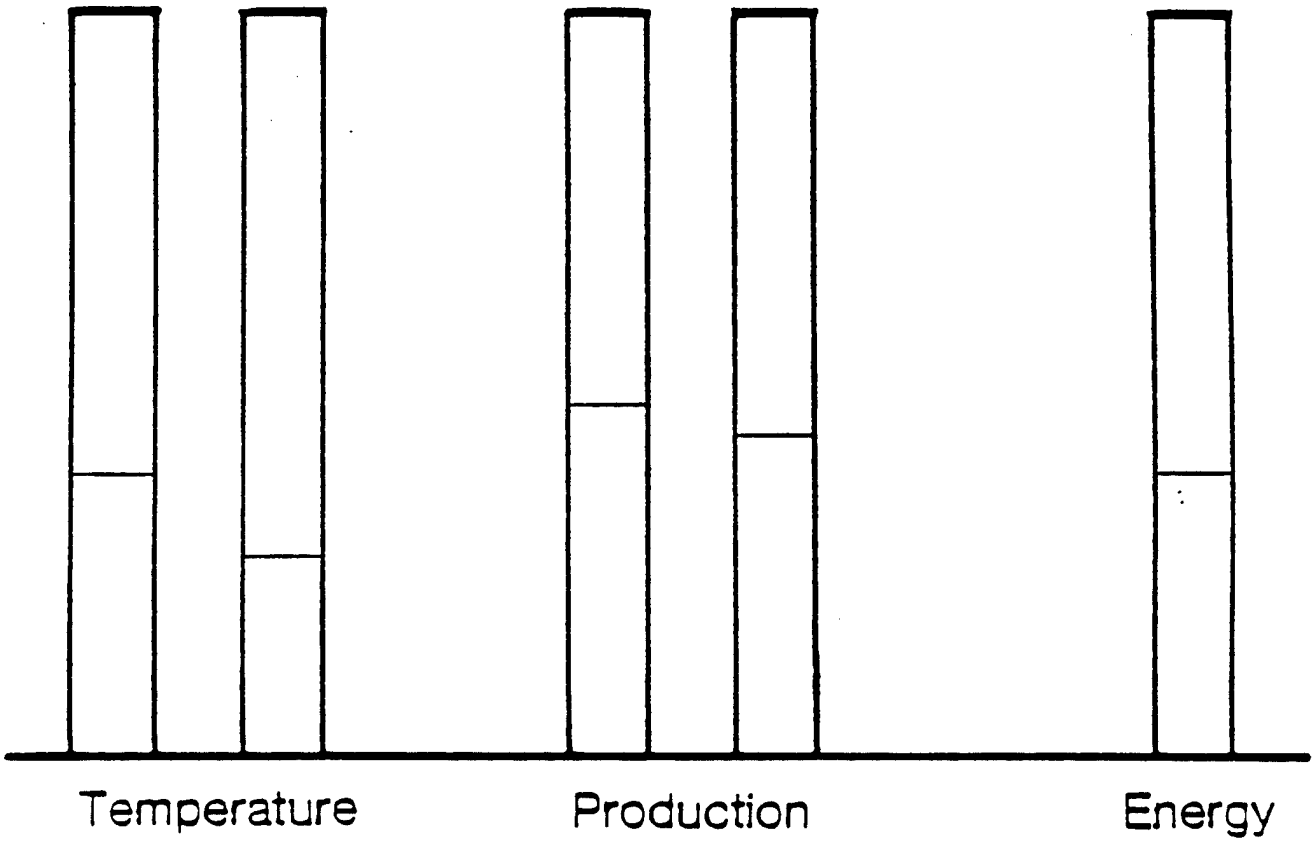


Forecast Data

Temperature	National Weather Service	36	°F
Temperature	Staff Meteorologist	25	°F
Production	Production Manager	45	units
Production	Quality Control Supervisor	41	units

End of Week Report

Week No. 31



Forecast Data

Temperature	National Weather Services	36 °F
Temperature	Staff Meteorologist	25 °F
Production	Production Manager	45 units
Production	Quality Control Supervisor	41 units

Actual Energy Consumption 362 units

RESULTS

The experiment is a 3 (groups) X 4 (blocks of 25 trials) factorial. Groups is a between variable, blocks is within. Two dependent variables, the lens model indices which measure the accuracy (r_a) and consistency (R_s) of prediction, were analyzed. The means and standard deviations are displayed in Table 1-2 (see also Figure 1-4). The Fisher's Z transforms of r_a and R_s (Z_a and Z_s) were then subjected to multivariate analyses of covariance, with the predictability of the environment (R_e) being the covariate for blocks. The analysis of covariance was called for since in the construction of the data set to be presented to the pilots, the correlations were controlled for the entire set of 104 trials, resulting in sampling variability across the four blocks. The sampling variability was of such a nature that the third block turned out to be an anomaly, with a high R_e and deviations from the desired cue-criterion correlations.

There was a highly significant Groups effect for accuracy (Z_a), with the MULTIPLE OBSERVATION subjects being more accurate. Even after the variation in R_e was covaried out, there was still a highly significant Blocks effect, but no Groups X Blocks interaction. The same pattern of significance obtained for consistency (Z_s) as for Z_a, but as Figure 1-4 shows, the subjects increased in consistency from Block 3 to Block 4 but decreased in the accuracy with which they predicted the environment (Z_a). In order to gain insight into the data without the potentially confounding third block, ANOVAs were conducted on the first two blocks alone. The only significant effect of interest was the expected groups effect.

Table 1-2

The means (M) and standard deviations (s) of Fisher's Z indices for each block for each group for Experiment 2.

Group		Blocks							
		1		2		3		4	
		Z_s	Z_a	Z_s	Z_a	Z_s	Z_a	Z_s	Z_a
Control	M	1.04	.62	1.16	.66	1.42	1.03	1.50	.87
	s	.32	.27	.25	.19	.23	.27	.26	.17
ERROR	M	1.16	.67	1.21	.67	1.63	1.05	1.77	.92
SHOWN	s	.28	.19	.44	.21	.50	.23	.35	.12
Mult. obs.	M	1.20	.86	1.26	.87	1.58	1.19	1.61	.92
	s	.29	.31	.38	.24	.26	.28	.36	.18

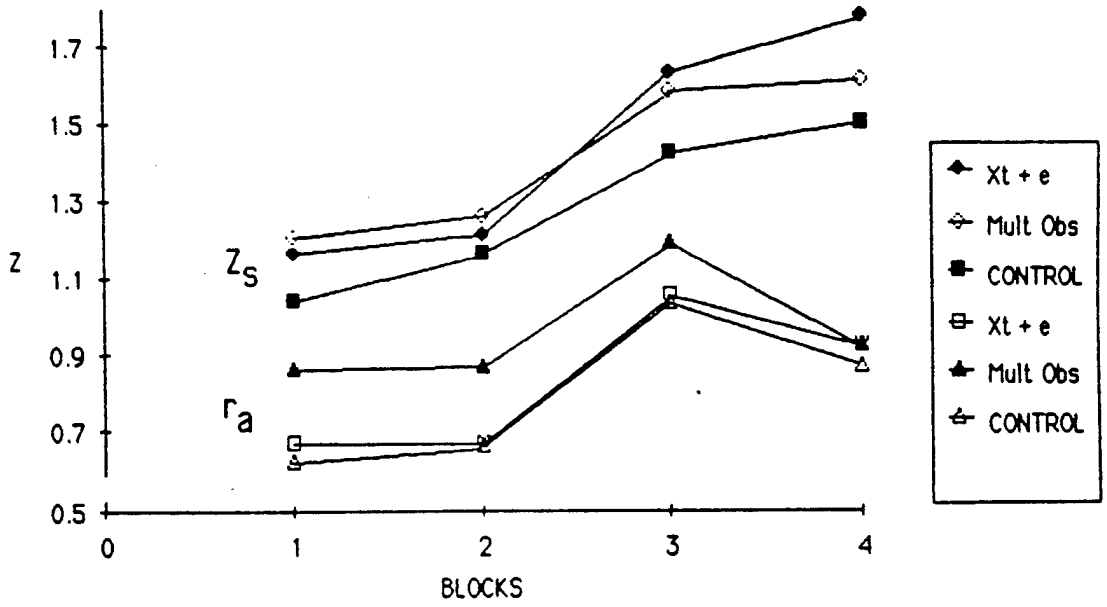


Figure 1-4. Consistency (Z_S) and Accuracy (r_a) for experiment 1.

Table 1-3

The results for the ANCOVA for the complete experiment for Z_a for Experiment 2.

Source	df	SS	F	p
Groups	2	.984	12.63	<.001
Blocks	3	.901	11.56	<.001
Groups X Blocks	6	.258	1.10	ns
Subjects X Groups	43	4.094	2.44	<.001
Error	129	5.028	-	-

Table 1-4

The results for the ANCOVA for the complete experiment for Z_s for Experiment 2..

Source	df	SS	F	p
Groups	2	.899	6.59	<.002
Blocks	3	8.035	46.79	<.001
Groups X Blocks	6	.296	.72	ns
Subjects X Groups	43	10.615	3.62	<.001
Error	129	8.798	-	-

Table 1-5

Results for the ANOVA for the first two blocks for Z_a for Experiment 2.

Source	df	SS	F	p
Groups	2	.985	12.06	<.001
Blocks	1	.004	.11	n.s.
Groups X Blocks	2	.010	.12	n.s.
Subjects X Groups	43	3.223	1.84	<.05
Error	43	1.755	-	-

Table 1-6

Results for the ANOVA for the first two blocks for Z_s for Experiment 2. .

Source	df	SS	F	p
Groups	2	.271	1.60	n.s.
Blocks	1	.146	1.72	n.s.
Groups X Blocks	2	.017	.10	n.s.
Subjects X Groups	43	6.016	1.65	n.s.
Error	43	1.755	-	-

DISCUSSION

The results for the S's ability to predict the environment are striking: the CONTROL and INPUT ERROR SHOWN groups are virtually identical, while the MULTIPLE OBSERVATION group starts out higher and stays higher as long as criterion feedback is provided. The drop in achievement on the test block (4) was unexpected, and we can only attribute that decrement to a

loss in motivation that seems to have occurred when the subjects were informed that feedback would no longer be provided after week 79. This conclusion is essentially speculation, and is based on observations by a number of introductory psychology students run at Bowling Green while the subjects of interest were being run at the Air Force base. Comparable observations by the pilots are not available.

In general, the results suggest that subjects are not disrupted by moderate amounts of ME error in the data on which their predictions are based. The MULTIPLE OBSERVATION subjects seem to have adopted some highly adaptive strategy to deal with the different values of the predictor variables, probably averaging, since their performance was quite good. In fact, the subjects with two observations of each predictor performed better than the theoretical upper bound of performance for the single observation conditions; they had to be doing something useful with the partly contradictory data, something akin to averaging over the random error. The data of the INPUT ERROR SHOWN pilots also suggests that the awareness of error in the predictors neither disrupted nor facilitated their ability to predict the criterion, compared to the CONTROLS.

EXPERIMENT 3

There are many ways of manipulating the awareness of ME error, and experiment 3 is a a computer implemented replication of the INPUT ERROR SHOWN condition of experiment 2, using college students as subjects. In this experiment, measurement error was manipulated by showing to the subjects the true values of the cues on each trial, after they had made their prediction for that trial. Subjects were assigned randomly to one of

four conditions, determined by two levels of each of two independent variables, task predictability (PREDICTABILITY: High vs. Low) and whether the subject was shown the locus of the input error (LOCUS: Shown vs. Hidden).

METHOD

Subjects. Sixty introductory psychology students served as subjects and received course credit for their participation.

Apparatus. An Apple MacIntosh computer was used to present stimuli and record subject's responses. Groups of 1-4 subjects were run simultaneously in a laboratory containing four computers. The room was designed so that each individual subject could not readily see any other subject's display screen. After brief instructions, the experimenter sat at a desk on one side of the room and was available to monitor and assist subjects throughout the experiment.

Procedure. Upon arriving, the subjects were seated in front of the computer which was displaying the appropriate cue display for the first trial. They read a brief set of printed instructions, asked questions if necessary, then proceeded with the task. The instructions described the same scenario as used in Experiment 2, that the experiment was a study in "how people use data", that they were to play the role of an executive trainee whose present task was to predict the weekly amount of energy consumed by the organization, to base their predictions on two forecasts, etc. On the monitor, the subjects saw two black bars representing the values of X_0, one labeled T (for temperature) and another labeled U (for units produced). A third black bar, identified by a number, provided the subjects a means of visually representing their predictions. The size of

these bars was scaled according to the MacIntosh screen, which uses "pixels" (a pixel being approximately .35 mm) as the units of measurement in the output window. At all times, windows on the screen reminded subjects of the meanings of the bars.

The entire screen is 491 pixels wide by 299 pixels high. Subjects saw a graph which measured 348 pixels horizontally and 200 pixels vertically. The bars, 12 pixels wide, were spaced equally on the horizontal axis. The height of the bars varied according to the cue values assigned to each bar.

Subjects made their predictions by moving the numbered bar up and down. When they pressed the key "U", the bar moved up; "D" moved the bar down. When they were satisfied that the bar represented their prediction, subjects depressed the return key. A fourth black bar, labeled E (for energy), appeared immediately on the screen and remained on for ten seconds. In the notation presented in the introduction, this bar represents F_0. The screen was then erased and the cue display for a new trial appeared. At this point in the procedure a manipulation, to be described below, was introduced. This procedure was followed for 104 trials, the first four of which were practice trials and dropped from the analysis. The entire sequence of trials took about 45 min to complete, after which a brief post-experimental questionnaire was administered.

PREDICTABILITY. The 104 trials were generated by combining values of temperature and units produced such that before error was introduced the criterion value was equal to $2/3\ X_{01} + 1/3\ X_{02}$. The multiple correlation between the cues and criterion, R_e, was manipulated by the magnitude of

with the Shown conditions, the following events occurred on each trial. Subjects first depressed the return key and the criterion bar appeared on the screen. Subjects were allowed to study the four bars on the screen indefinitely. After ten seconds, the computer beeped and a message appeared at the top of the screen indicating that pressing the "P" key allowed them to proceed to the next trial.

RESULTS

Multiple regression analyses were run on each subject's data and the len's model indices R_s (consistency) and r_a (accuracy) were obtained for each subject, as in Experiment 2. These indices, after Fisher's Z transformations, were treated as dependent variables in a multivariate ANOVA, with Predictability (high vs. low) and Locus (Hidden vs. Shown) of error as the independent variables. Subjects in the High task predictability condition did significantly better than those in the Low task predictability condition with respect to both R_s ($F = 7.76$, $p < .007$) and r_a ($F = 27.80$, $p < .0001$).

Showing subjects that there was measurement error in the input side of the environment had no effect on either their consistency ($F = 0.17$, n.s.) or achievement ($F = 0.01$, n.s.). Nor was there a significant interaction between Level and Locus of error ($F = 0.03$ for R_s and $F = 0.04$ for r_a).

Trials 51-100. Performance in an MCPL tasks tends to improve rapidly during the beginnning of the task and then level off. Hence, the second half of the trials were analyzed separately. In addition, these last 50 trials were split into two blocks of 25 and this blocks variable was then treated as a within subjects factor in a repeated measures analysis of variance.

Again, R_s and r_a were the dependent variables and error Locus and task Predictability were the independent variables. Results were the same as when the entire set of trials was analyzed. A main effect for Level of error was obtained for both R_s ($F = 14.41$, $p < .0004$) and r_a ($F = 35.91$, $p < .0001$). However, there was no effect for Locus and no interaction between Predictability and Locus.

Univariate tests of the within subjects factors revealed no main effect for blocks, which indicates that subjects' performance remained relatively stable between trials 51-75 and 76-100. There was no interaction between blocks and Locus, but there was an interaction between blocks and Predictability for r_a ($F = 50.90$, $p < .0001$). This interaction was not significant for R_s, however. Finally, there was no significant three-way interaction between block, Locus, and Level.

DISCUSSION

This finding of highly significant Predictability effects, typical in MCPL tasks, confirms that the subjects were attending to the task. The failure to find significant effects for Locus, with the sample sizes involved, suggests that knowledge of the presence of ME error in the input data does not influence the prediction strategies of subjects to an interesting degree, if at all. This is tantamount to saying that subjects behave optimally with regard to the processing of measurement error in the information on which they base their predictions, since the rational strategy in this environment is to ignore the error and to attempt to compute the correlations between X_0 and F_0. The conclusions of this study, then, are consonant with those of Experiments 1 and 2.

EXPERIMENT 4

The purpose of this experiment is to investigate the possible effects of making subjects aware that there is measurement error in the feedback. It has been generally accepted that learning in a standard MCPL environment is highly inefficient, and that one possible reason is that subjects "chase error" in the feedback; in effect subjects see the task as a deterministic one rather than as a probabilistic one (Brehmer, 1980). Thus subjects may perceive the task as requiring them to be exactly correct on each trial. This leads us to the somewhat paradoxical expectation that, should we give subjects less precise feedback, then they might not see the task as deterministic, and might not futilely chase random error.

In this experiment, the performance of subjects given the exact value of F_0 will serve as a baseline against which to assess the performance of subjects who are given feedback about the range of values into which F_0 would fall 2/3 of the time. If the above speculation about the assumption by subjects that the MCPL task is deterministic is correct, then the performance of the group getting the nonspecific feedback should exceed that which gets the typical point feedback.

METHOD

Subjects. Forty introductory psychology students served as subjects as part of a course requirement.

Apparatus. The apparatus and procedure were the same as in Experiment 3, except as noted.

Procedure. Subjects were assigned randomly to one of two conditions, a standard feedback condition, F_0, and a condition in which the feedback was

presented as a range, F_0 + e. Subjects were seated in front of a computer and instructed as above. On the computer screen, subjects saw two hollow columns representing the input values, X_i, labeled T and U, and a third, hollow column for predictions (see p. 60), labeled with the subject's initials. The dimensions of the screen and graph were identical to experiment 3. The only exception is that all three columns were of equal height (200 pixels). Each hollow bar contained a prominent, horizontal line somewhere across its width. The lines in the first two columns represented the cue levels. The use of columns was a departure from typical presentation format, and was adopted to facilitate the representation of F_0 + e.

Subjects again moved the prediction bar up or down by pressing either "U" or "D" on the computer keyboard. After making their predictions, subjects pressed the return key and a fourth hollow column, E, appeared on the screen. The representation of F_0 in this column constituted the independent variable in this study, as described below. All four bars remained on the screen indefinitely and subjects were allowed to study them for as long as necessary. When finished, subjects pressed "P" and the screen was cleared for a new trial.

Subjects followed this procedure for 104 trials, of which the first four were practice and thus not analyzed. The relationship between the cues and criterion was identical to that used in Experiment 3. Only one level of error was used, R_e being equal to .71 in both conditions. Cue validities were the same as in experiment 3, .78 and .40. Again, either T or U could be

PRESS U TO MOVE BAR UP;
PRESS D TO MOVE BAR DOWN;
PRESS <RETURN> TO SEE PREDICTED VALUE OF E

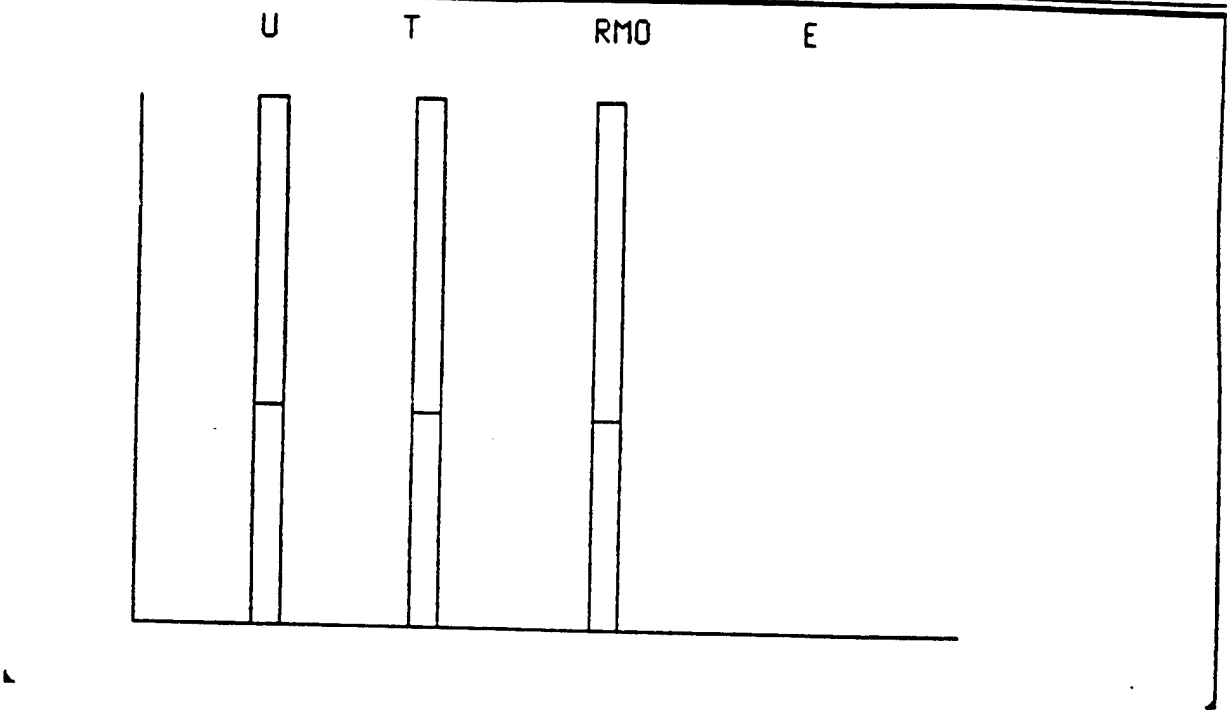


Figure 1-4. The input display for experiment 4.

the more valid cue, to avoid confounding cue validity with cue label.

After making their prediction on a given trial, subjects pressed the return key and a fourth hollow bar, the feedback bar labeled E, appeared on the screen. In the F_0 only condition, subjects saw a straight line inside the bar (see p. 62) which indicated the point prediction of the value of energy consumption, as contaminated by an error component as in a standard MCPL study. In the F_0 + e condition, subjects saw a shaded portion of the otherwise hollow column, and were told that the actual value of the criterion would be inside this range on 2/3 of the trials (see p.63). This range was generated by adding and subtracting one standard ~~standard~~ error of estimate from the criterion value after error was added to it.

RESULTS

As in Experiment 3, the trials were split into blocks of 25 and Blocks treated as a within subjects factor. A repeated measures ANOVA was run with R_s and r_a as the dependent variables, feedback type (F_0 only vs. F_0 + e) as the between groups factor and Blocks as the within groups factor on blocks 3 and 4. There was absolutely no main effect for feedback type in either R_s ($F = 0.00$, n.s.) or r_a ($F = 0.10$, n.s.).

Univariate tests of the within subjects factors revealed a significant main effect for Blocks in R_s ($F = 9.78$, $p < .0034$) and r_a ($F = 11.33$, $p < .0018$), indicating that subjects did better in the last block of 25 trials than in the third block of 25 trials. No Blocks X feedback type interaction obtained for R_s, but this interaction was significant for r_a ($F = 5.92$, $p < .0198$).

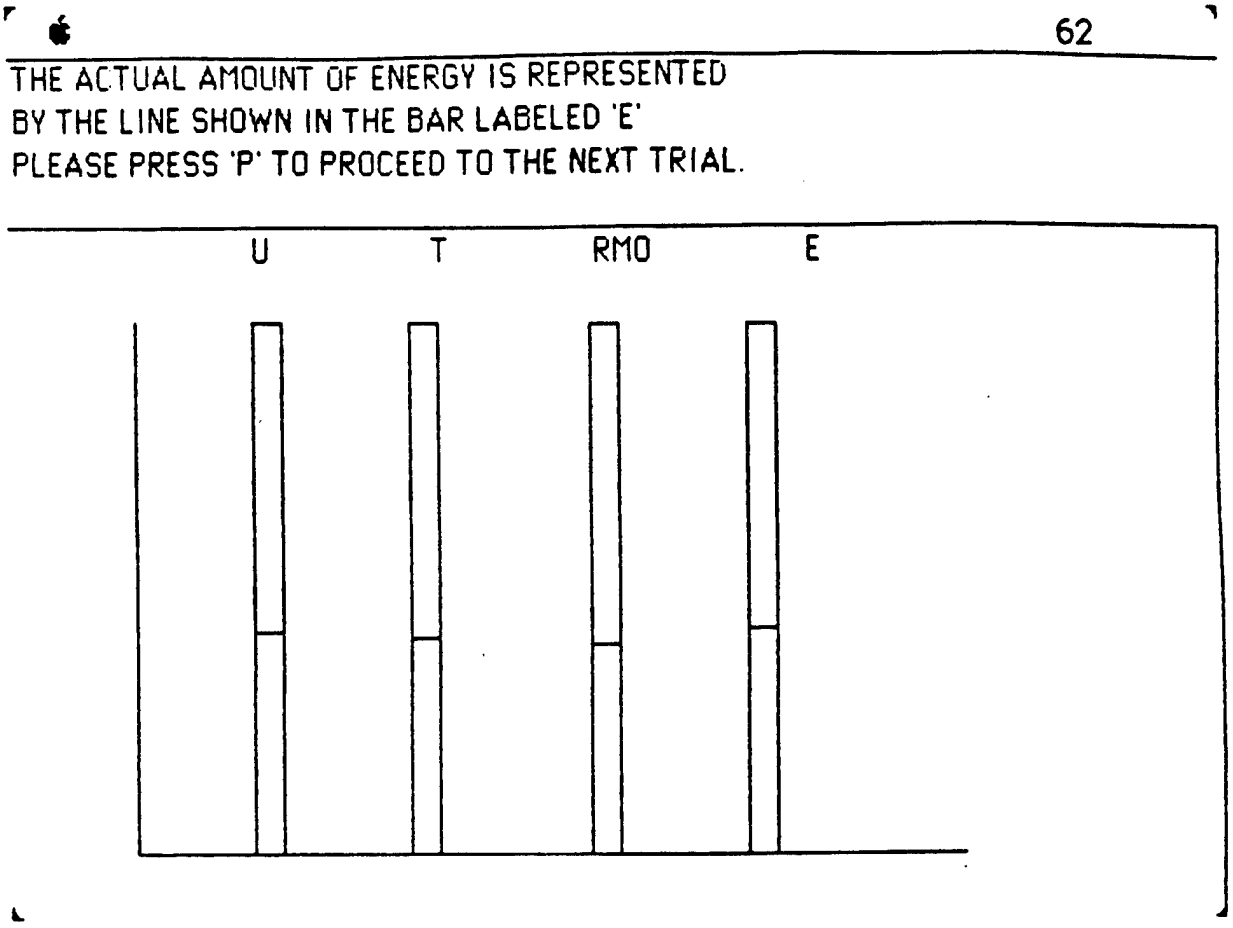


Figure 1-6. The feedback display from the F_0 condition of experiment 4.

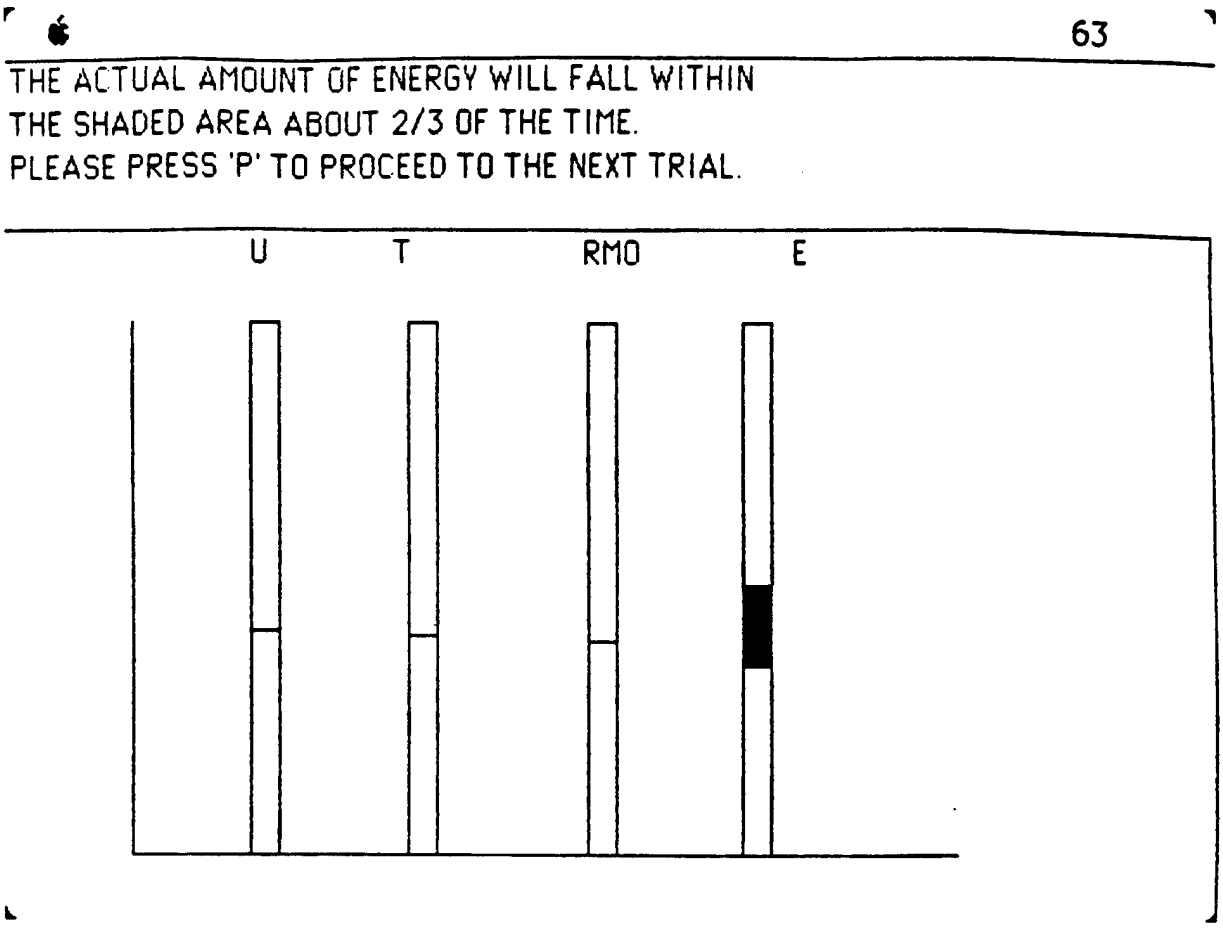


Figure 1-7. The feedback display from the F_{o+e} condition of experiment 4.

DISCUSSION

It appears that indicating to subjects that error is present in the feedback side of the environment makes as little difference as showing them it is in the input side. This was unexpected.

EXPERIMENT 5

This study employed the MCPL paradigm to investigate the effect of error type, i.e., ME vs. SF. No distinction is made concerning the locus of error in this study, that is, even though subjects might assume that there are errors on some trials, nothing in the procedure to be described would allow the subjects to discern whether that error was in the X_0 or in the F_0.

METHOD

Subjects. Forty introductory psychology students served individually as subjects in partial fulfillment of a course requirement.

Apparatus. Stimulus presentation and response recording was accomplished on an Apple II+ computer, with a 16 in Sanyo monitor on a stand directly above the keyboard, approximately at eye level. The subject was seated at a comfortable, self-adjusted, distance from the apparatus.

Procedure. Subjects were assigned to one of two treatments, ME or SF Error. Both verbal and written instructions were given. First the subject typed in his or her initials. Immediately three vertical bars appeared on the monitor, labeled A, B and with the subject's initials, from left to right. The displays were similar to those shown above. The subjects predicted the value of a target, C, from the values of A and B. They were told (in simple terms) that the function forms relating C to A and B were both positive linear, and instructed how to respond. Subjects made their

predictions by moving the vertical bar which was labeled with their initials. One key controlled upward movement, another controlled downward movement. When the subject was satisfied that the bar represented the prediction, he or she pressed "return". Upon depression of return, a fourth bar, labeled C, appeared and remained on for about 4 seconds. The screen was erased, and a new trial with new values of A and B appeared. The initial height of the response bar was selected randomly on each of 100 trials. In the notation of Fig. 1, A and B are X_{0i} values, the final height of the response bar is Y_s, C is F_0.

The ME condition. A and B could take on any value from 1 to 10. The 100 trials were composed of a factorial combination of A and B, such that before error was introduced the criterion value was the arithmetic average of A and B. The variance of e determined the multiple correlation between the cues (A and B) and the criterion (C), R_e. The computer added the random component to the sum of A and B before presenting the C bar to the subject.

The SF condition. The same factorial combination of A and B served as cues. In this treatment, however, Y_e was exactly the mean of A + B, except that on some predetermined number of trials a random value was selected, with replacement, from the distribution of (A + B)/2 and presented as Y_e. Thus, in the SF condition, a trial with A = 1 and B = 1 would be, on the average, predictive of a very low value of C, but on an SF trial might be associated with a C of 10. In both ME and SF, error levels were selected randomly for each subject, given the restriction that moderate to high levels of R_e should result.

Twenty four subjects were run in the SF, 16 in the ME condition. At the time the study was conducted we knew of no analytical means by which a given R_e in the SF condition could be predetermined, a problem we solved, at least in large part, later (Doherty & Sullivan, in press).

For the ME condition, the R_e could be determined ahead of time by selecting an appropriate variance for the error term to be added to either side of the prediction equation. Such constants were sampled randomly from a distribution of values that would give relatively high R_e values. Eight subjects in the SF condition for whom $R_e < .60$ were excluded from analyses involving comparisons with the Measurement Error group, leaving 16 subjects per group. The mean R_e for the remaining 16 SF Ss was .81, while the corresponding mean in the Measurement Error group was .87. The difference between mean R_e values was not significant ($t(30) = 1.39$, n.s.). Note that this operation loaded the dice against the SF condition.

RESULTS

In order to assess the impact of error type on predictions a MANOVA was run with SF vs. ME as the independent variables, and with R_s and r_a as the dependent variables. Both R_s and r_a were higher for System Failure (contrary to the direction of difference for R_e) but the overall test was not significant (for R_s $p<.07$, for r_a $p<.13$).

Trials 51-100. The second half of the trials was analyzed separately. The mean values of Re were again similar, with R_e for SF and ME being .84 and .88, respectively. The MANOVA on r_a and R_s was significant ($p < .01$),

the difference being due to the differences in R_s, .95 vs. .84, for SF vs. ME, respectively. The r_a values did not differ, being .76 and .70, respectively.

An Analysis of the SF data only. The error to which SF subjects were exposed was of such a nature that they could often be exactly correct, were they to employ an analytical approach (Hammond, 1986) and to consider the trials on which spurious feedback occurred as irrelevant. Thus, an analysis of the data of all 24 SF subjects was conducted. The mean value over all trials of the correlation between Y_s and Y'_e (see figure 3 for the relation obetween these terms) was significantly greater than R_e, via a t-test for correlated observations, and numerically greater than R_e for 21 of the 24 subjects. But the theoretical upper bound of the correlation between Y_s and Y'_e is R_e, given a linear environment. Thus, subjects had developed a model of the environment that was superior to that normally considered possible, that is, they apparently learned to ignore the error trials and to maintain the correct model which had been deliberately designed to be as simple and obvious as possible.

DISCUSSION

The manipulation of the SF error in this study, as in Kern (1982), had a significant impact on judgments, though in the opposite direction from that found by Kern. Many subjects were able to ignore error trials, and to continue to predict the criterion accurately. This is especially clearly shown in the high correlation between Y_s and Y'_e. Clearly, error type has powerful effects on behavior, but these effects have as yet unknown situational determinants.

The study was one of the earliest we did, and was an early test of the potential effects of the difference between SF and ME error. Given such a limited goal, it was designed with the barest of demands on subjects' learning abilities, i.e., equal weighting of variables with positive linear functional relations to the criterion. There are, however, several generalizations which can be asserted based in part on the data, in part on the larger body of MCPL research. First, note that the SF subjects did extraordinarily well on what might have been expected to be a truly difficult task. One explanation is that the simple task allowed, or even elicited, analytical rather than intuitive thought (Hammond, 1986). But the subjects' rather impressive performance in the face of, in some cases, so many error trials, requires further exploration.

These results strongly suggest a hypothesis that , if true, should have a significant impact on the the design and analysis of MCPL studies, and on some aspects of our conception of feedback systems. That is, subjects are powerfully rewarded, and presumptively equally powerfully influenced by, a "direct hit," a trial on which Ys-Ye = 0. There is an implication in these results that the typical subject has a criterion for error which is radically nonlinear with Ys-Ye. Nonlinearity is implicitly assumed by investigators who use root mean squared error (RMS) in designing feedback systems. But RMS nonlinearity is such that the penalty per unit error is greater as Ys-Ye increases. The present results and observations of subjects' behavior suggest that the the largest psychological penalty occurs with the subject's observation that Ys-Ye = 0. This speculation is treated in another context in Doherty & Balzer (in press).

Experiment 6

The present study employed a manipulation using a plant growth task similar to the one used by York et al. (1987) in order to determine the differential effects of SF and ME error. In one condition (ME), the two cues were each degraded by ME error only. In a second condition (SF), subjects also encountered, on 30% of the trials, SF error in one of the two cues. It was anticipated that subjects presented with ME error would do better at the task than subjects confronted with both ME and SF error, even though the SF task was constructed to have a more predictable environment.

Method

Subjects. Thirty introductory psychology students participated in groups of 1-4 in partial fulfillment of course requirements.

Materials. Macintosh computers were used to display two "meters", one labeled "Level of Water" and one labeled "Level of Fertilizer", using a 40 point graduated scale marked in increments of 5 (see page 70). To avoid confounding cue label with cue validity, cue label was randomized within experimental conditions, so that cue 1, on the left, was water about half the time and fertilizer about half the time. On each trial, subjects were asked to predict a plant's growth based on the two cues. In pilot testing most subjects reported that they had relied heavily on their previous knowledge of how water and fertilizer affect plant growth, hence the instructions stressed that previous knowledge would not benefit them on this task and should be ignored. The subject's response was displayed immediately to the right of the cues, and, to the right of that, "Actual Growth" was given as outcome feedback on a scale marked in increments of 1 from 1 to 9 (see p. 71). The units chosen for the cue and criterion scales

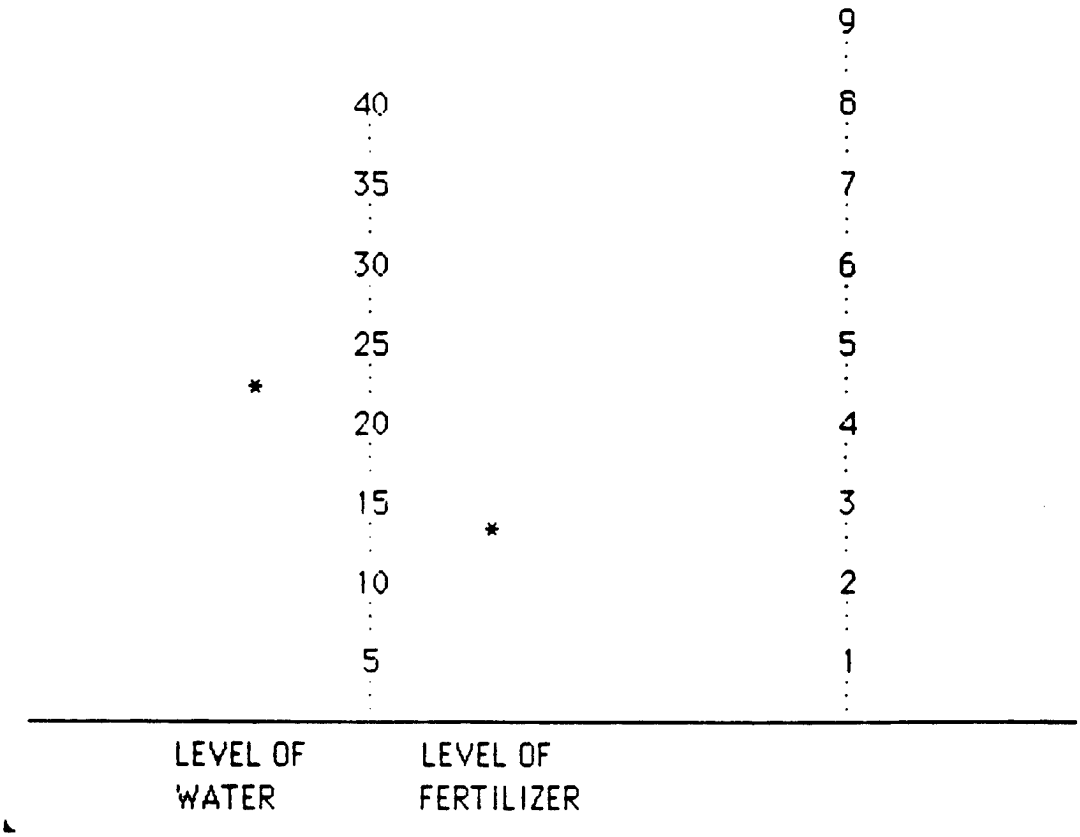


Figure 1-8. The input display for experiment 6.

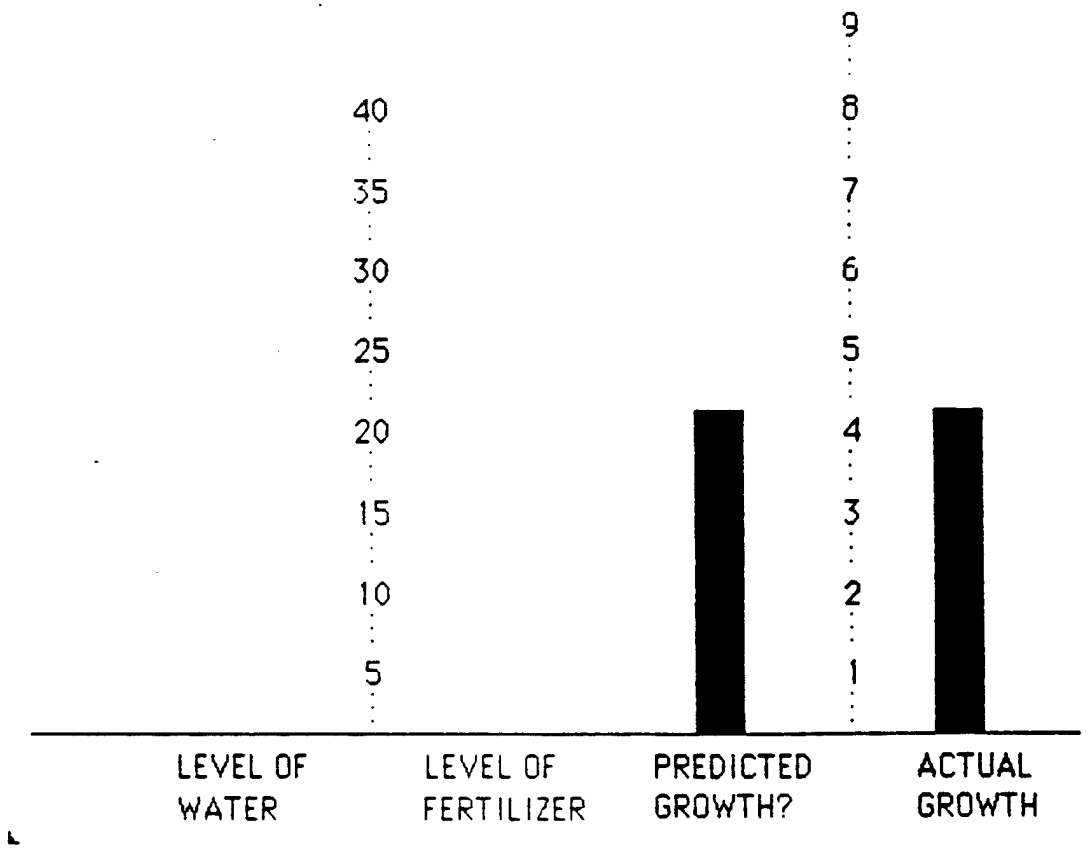


Figure 1-9. The feedback display for experiment 6.

were arbitrary.

Task Design. There were 52 two-cue trials, with outcome feedback given on each trial. Each cue was allowed to take on true values ranging from 2 to 38. The cues were uncorrelated, and had positive linear relationships with the criterion. Error in the cues was operationalized by treating each cue as a true value and adding random error to the true value to produce observed values. True score cue validities for the ME treatment were .81 and .32, for cues 1 and 2, respectively, with a total task predictability (R_e) of .87. For the SF treatment the R_e based on the true scores was .99, and the true score validities of the cues were .89 and .46. Five observed values were generated for each cue on each trial, and were given to subjects sequentially as multiple meter readings. These were represented as asterisks adjacent to the graduated scale.

For the ME condition, the error was a random variable sampled from a population with a mean of 0 and a standard deviation of 2. On each trial five error values were added to the true value to produce five observations. For the SF condition, a trial was designated an SF trial with a probability (p) of .3. The five observed values for cue 1 on an SF trial consisted of values chosen randomly, with replacement, from the distribution of possible cue values, without regard for the functional relationships between cue and criterion. Thus there were about 15 (.3 x 52) such trials. Only cue 1 contained system failure error on these 15 trials; cue 2 was generated as in the ME only condition. On the remaining approximately 37 trials, both cues had ME error only.

A number of SF error distributions were generated according to the above statistical characteristics, and, in light of our expectation that SF

error would be more disruptive than ME error, 15 were chosen so that the mean task predictability based on observed values for the SF condition would be greater than for the ME condition. The average R_e based on all observed values, i.e., after error was added to the cues, was .90 for the ME treatment and .92 for the SF treatment. The difference between the means of the R_e values in the two treatments was significant ($t(28)=3.05$, $p<.01$).

Procedure. Four computers were arranged in a large laboratory so that subjects could not see one another while they were performing the task. Upon arriving, each was seated at a computer and provided with printed instructions. Time was allowed for subjects to ask any questions or voice any concerns about the task.

On each trial, the first observations of cues 1 and 2 were presented simultaneously for about one sec, followed by the simultaneous presentation of the second observed values, and so on. Each pair of cue values was erased prior to the presentation of the next pair. In short, on each trial, subjects saw five pairs of cue values flash on the screen during a 5-sec interval. After the fifth pair of cue values was erased, subjects were prompted by the computer to predict the growth of the plant for that trial. Subjects entered their predictions by depressing a single numeric key (ranging from 1 to 9) on the computer keyboard. Subjects had as much time as they wanted to make their predictions. Immediately after subjects entered their predictions into the computer, "Actual Growth" (i.e., the criterion) was given as outcome feedback. This remained on the screen for about ten sec, after which the screen was cleared and the next trial began. Upon completion of the 52 trials, subjects filled out a post-experimental questionnaire. This included questions about which predictor they would

use if they could only use one to predict from and the "percentage of importance" they thought each predictor had. Subjects were then debriefed and thanked for their participation.

Results. The basic index of performance in this investigation was the accuracy with which the subjects predicted the criterion value, r_a. The mean r_a values are presented in Table 1.

Table 1

Mean lens model indices[1] for all trials and for the subset of trials.

	All trials			Subset of trials		
Condition	r_a	R_{S1}	R_{S2}	r_a	β_2	SD
ME	66	78	78	70	27	1.86
SF	53	59	68	07	24	1.50

[1]Mean correlations are via Fisher's Z transformation, decimals omitted. The two multiple R values differ with respect to what the responses were regressed upon: R_{S1} on the true scores; R_{S2} on the true scores except that on SF trials the mean of the five SF observations replaced the true scores displaced by those errors. β_2 is the beta weight for the second cue, i.e., the cue without SF error.

As anticipated, subjects in the ME condition had a significantly higher r_a than subjects in the SF condition ($t(28) = 2.52$, $p < .02$). In addition, the variance of r_a for the SF group was greater than the variance for the ME group ($F_{max} = 5.80$, $p < .01$). While the assumption of homogeneity of

variance was not met, the fact that there were an equal number of subjects in both conditions makes the difference between the means interpretable (Kirk, 1982).

The multiple correlation between the cue values and the subject's predictions (R_S) is normally a measure of how consistently subjects used a linear policy in making predictions, but in this study the multiple observations make R_S a much less straightforward index of consistency than in a traditional MCPL study. One value for R_S (call it R_{S1}) was obtained by regressing the subject's predictions on the true values of the cues. Subjects in the ME condition had a significantly higher R_{S1} than subjects in the SF condition ($t(28) = 3.66$, $p < .01$). Regressing the subject's judgments on the true scores in this study is not fully appropriate, however, since in the SF error trials subjects never saw the true values, while in the ME condition the mean of the five observations on a trial was approximately equal to the true score on that trial. For any subject who averages the five cues and uses the average as a basis for prediction, consistency in cue usage will be underestimated by R_S, especially in the SF condition. In an effort to overcome this problem, we went back to the SF trials and replaced the true score values for cue 1 with the mean of the five random observations on that trial. This value is a more accurate representation of the value that the subject who was trying to use cue 1 might infer on the SF trials. This value, R_{S2}, was not significantly higher for subjects in the ME condition ($t(28) = 1.89$, n.s.).

In order to determine how subjects were using the cues, one would

normally look at the beta weights from the regression analyses. However, as should be clear from the discussion of R_S, the beta weights calculated on all 52 trials would provide little insight into why subjects in the SF error condition did so much worse than subjects in the ME condition, because SF error contaminates the regression analyses used to calculate these weights. Therefore, the data were broken up into subsets.

Analyses on Subsets of the Data. One might speculate that erratic readings introduced by SF error may have served as a "signal" to the subjects to ignore cue 1 and concentrate on cue 2 in making their predictions. This was tested by conducting regression analyses on only those trials which had SF error and comparing the results to regressions done on a random sample of an equal number of trials from the ME condition subjects. A significant difference in the appropriate direction between the betas for cue 2 (β_2) between the two groups would indicate that subjects in the SF condition had indeed used the high within trial variance of the observations of cue 1 as a higher order cue, and were simply ignoring the cue with SF error on those trials. The results of these analyses are presented in Table 1. This hypothesis was not supported; β_2 for the ME group were not significantly higher than those for the SF group; ($t(28) = .23$, n.s.). Had subjects ignored cue 1 on these particular trials, they could have had much higher achievement scores, as indicated by the correlation between cue 2 and the criterion. This is, of course, true of both conditions, even across all 52 trials, but the r_a values on the 15 SF trials were extremely low.

Normally, one thinks of the regressiveness of responses as an index of

whether subjects had properly accounted for cue unreliability (Kahneman & Tversky, 1973). Hence the standard deviations of the subjects' responses on the subset of the trials in the two conditions were calculated. The SF standard deviations were lower than those of the ME subjects ($t(28) = -2.62$, $p < .02$). In these circumstances, however, the presence of regressiveness of responses is not sufficient evidence to infer that subjects were appropriately discounting the unreliable data. Consider an SF subject who is averaging and using the 5 observations. The low variance of the means of those 5 scores across trials and the tendency of those means to cluster around the midpoint of the scale would lead to predictions with less variance. Hence in this case a diminished response variance might well be reflecting insensitivity to cue unreliability.

Subjective Weights. The subjects were asked to state which cue, water or fertilizer, they would use if they could only use one, and to distribute 100 points between the two cues. An inspection of the subjective weights indicated that subjects were able to discern the environmental structure; 27 of the 30 subjects correctly chose the more important cue. For 26 of the 27, the subjective weights were in the same order as the ecological validities. One subject selected the correct cue but gave equal weights.

Discussion

The results indicate that, unlike ME error, SF error significantly reduces an individual's ability to cope with an uncertain environment. Subjects who were presented with the classical form of error (ME) did reasonably well on the task, in terms of r_a, though nowhere close to the

statistical limit of achievement. But the presence of error in the form of random observations unrelated to the data generating process, i.e., SF error, degraded performance still further. This is in spite of the fact that the R_e for the SF group was significantly (though only slightly) higher than for the ME condition. In essence, we deliberately "stacked the deck" against finding a significant difference, since the criterion was more predictable in the SF condition. However, subjects in this condition still performed poorly.

A possible strategy for subjects to deal with the perceptually salient uncertainty introduced by SF error was alluded to earlier. Subjects could have ignored the cue with SF error (cue 1) on those trials. They could have focused on and used only cue 2 (the less valid cue) for their predictions. This hypothesis was tested and found to be untenable by the failure to find a difference between β_2 values in the subset analysis. It appears that subjects were attempting to use the cue with SF error in it, although it is unclear just how they were using it.

While the subjects in the SF condition did not perform well in the sense of being able to predict the point values of the criterion value accurately, they were surprisingly accurate in making the dichotomous choice of which cue was more important. SF error had a much less disruptive effect on their retrospective accuracy in cue selection than we had anticipated.

SF errors may cause disruption in other paradigms as well, though "error" must be defined differently in these tasks. We now turn our attention to a second paradigm in which we have manipulated error, Wason's 2-4-6 task.

PART 2

RESEARCH ON DATA ERROR USING WASON'S 2-4-6 PARADIGM

A. Prior Research

B. Research Conducted Under This Contract

Experiment 7. Single-Rule and Two-Rule versions by three Feedback conditions: 1) No Error, 2) Informed Error, and 3) Uninformed error

Experiment 8. The effects of Informing vs. Not Informing on Error and No Error conditions- Single-Rule version only.

A. PRIOR RESEARCH

According to Popper (1959), rational scientific inquiry should involve active attempts to disconfirm, rather than confirm, proposed hypotheses. However, many studies of hypothesis testing using error-free data have shown that most subjects prefer to examine evidence predicted by the hypothesis to occur, a "positive test (+test) strategy" (Klayman and Ha, 1987, p. 213). Such data are most likely confirm, and cannot disconfirm hypotheses (Wason, 1960; Mynatt, Doherty & Tweney, 1977, 1978; Tweney, et al., 1980). Such seemingly irrational behavior, shown as well by many working scientists (Mitroff, 1974), may serve as a heuristic for developing a hypothesis and establishing the reliability of the data before attempting to disconfirm (Mynatt, Doherty & Tweney, 1977, 1978; Tweney, Doherty & Mynatt, 1981; Klayman & Ha, 1984; Tweney, 1985).

Gorman (1986) used a group problem-solving task based on a card game called "Eleusis" to assess how warning subjects that system failure error might occur affected hypothesis testing. The Eleusis task involved having subjects play individual cards to discover sequencing rules about card order, such as 'Alternating red and black'. Groups of four subjects were assigned to one of three types of strategy instructions: confirmatory (emphasizing +tests), disconfirmatory (emphasizing -tests), or no strategy. In an earlier study using Eleusis, Gorman et al. (1984) had demonstrated that groups shown how to disconfirm their hypotheses using a negative test strategy performed significantly better than groups either shown how to confirm hypotheses using a positive test strategy or given no strategy instructions. All groups were told that "On anywhere from 0 to 20 per cent of the trials, the feedback you receive will be inaccurate"

(Gorman, 1986, p. 89), though none of the feedback actually contained error. Gorman (1986) found that knowing that the data might contain error severely disrupted performance on the task, even for groups given disconfirmatory instructions, since a significant amount of subjects' time was spent replicating tests to check for error. In addition, subjects appeared to use the knowledge of the possibility of error to classify potentially useful disconfirmatory evidence as error, which also disrupted performance.

While Gorman found that subjects who knew about the presence or possibility of system failure error tended to ignore disconfirming data and/or preferred to replicate only disconfirming trials, several methodological difficulties were present which could have affected interpretation of the results. Gorman focused on group problem-solving, used a possible (0-20%) rather than an absolute error information condition, and did not include an actual data error condition and a no-error instructional condition for comparison to the possible error conditions.

Wason's 2-4-6 Rule Discovery Problem. Based in part on the results of studies by Gorman and by Kern (1982, cf. infra), as well as on the issues raised by these studies' methodological limitations, a major purpose of the studies to be reported was to compare the effects on hypothesis-testing heuristics of providing vs. not providing information that error might occur by utilizing a well-documented experimental task under both actual error and no error feedback conditions. The experimental task chosen, Wason's (1960) 2-4-6 rule discovery problem, has frequently been used to evaluate the roles of positive (potentially confirmatory) and negative (potentially disconfirmatory) test strategies involved in hypothesis testing under various instructional conditions (Tweney et al.,

1980; Walker & Tweney, 1983; Walker, 1985, 1986; Gorman & Gorman, 1984; Klayman & Ha, 1985; Tukey, 1986).

In Wason's original version of the task, subjects attempted to discover the general number-sequencing rule, "three ascending numbers," when given the sequence, "2, 4, 6", as a positive instance of the rule. They could test their ideas about the rule with additional sequences, which the experimenter responded to as either fitting or not fitting the general rule. Subjects were asked to state the rule when they felt sure they knew what it was, based on the outcomes of their tests. If the first rule announcement was wrong, they could continue to make more tests and rule announcements. Wason's choice of a very general rule and a misleadingly specific example was intentional, "...so that several plausible hypotheses about an unknown rule could be supported by citing instances which confirmed them, or refuted by citing instances which disconfirmed them" (Wason, 1962, p.250). For instance, a plausible hypothesis (e.g. "three even numbers") could only be disconfirmed by using a negative test strategy (e.g. testing a sequence inconsistent with the hypothesis, such as "1-2-3"), but which turned out to be consistent with the experimenter's rule.

The utility of a negative test strategy was demonstrated by the small number of subjects who correctly stated the rule on the first announcement. These subjects "tended both to eliminate more possibilities, and to generate more negative instances than did those who announced a first incorrect rule" (Wason, 1960, p. 139). On the other hand, of those subjects who first announced an incorrect rule, the majority initially proposed one very specific hypothesis, such as "three even numbers" or "numbers separated by two". Generally these subjects gathered confirming evidence for their hypothesis by conducting positive

tests, such as "even numbers--6, 8, 10", and did not attempt disconfirmation by conducting negative tests, such as "even numbers--3, 5, 7". Thus, by focusing their attention on positive instances of their hypotheses, most subjects declared rules that were sufficient to explain the obtained evidence, but did not specify both the sufficient and necessary conditions for all the possible evidence that could actually fit the rule.

Error Studies Utilizing the 2-4-6 Problem. In most studies of the Wason 2-4-6 task, subjects were given accurate feedback about whether or not a hypothesis test fit the rule or not, though two studies have included error or the possibility of error. Markowitz and Mynatt (1982), using a computer-implemented version of Wason's task with and without system failure error, reported that, of five subjects given some erroneous feedback, all were much less likely to give up proposed hypotheses and more likely to retest data if the feedback disconfirmed their current hypotheses. However, the sample was extremely small and, as in Kern's (1982) study, Markowitz and Mynatt did not differentiate between the effects of potential error information and actual system failure error on hypothesis-testing heuristics.

Gorman (personal communication, February 10, 1987) recently completed a series of four experiments utilizing progressively more difficult versions of the Wason 2-4-6 task to study the psychological effects of the knowledge of the possibility of system failure error when no error was actually present. As in Gorman's 1986 Eleusis study, subjects in the possible error conditions were always told that 0 to 20 per cent of the trials might contain random feedback error. The first experiment, comparing possible error to a no possible error condition, did

not find significant differences in solving rates, though possible error subjects used significantly more trials to check for errors before reaching the correct solution. In the second experiment, two dimensions, color and letter, were added as distractions to the task. Subjects wrote each sequence in either red or black or a combination of red and black followed by one of 26 letters of the alphabet. They solved for an alternating sequencing rule (either odd-even-odd or even-odd-even) under possible error and no possible error conditions. Only a few subjects were able to solve the task and there were no significant differences between conditions in solution rates. As in the first experiment, possible error subjects used significantly more trials, as well as significantly more repeated trials and trials they expected to be incorrect.

Since the results of Gorman's second experiment appeared contrary to those obtained earlier for the Eleusis task (Gorman, 1986), a third experiment was conducted in which the alternating rule was used again without the added color and letter dimensions. All subjects were initially given ten sequences and the results of ten tests under either possible error or no error conditions. After reviewing the sequences and tests, subjects were allowed to test five additional sequences. Providing initial sequences and limiting the number of tests was designed to inhibit subjects' use of repeated trials. However no significant differences in solution rates, expected number of incorrect trials, and number of repeated trials were found between the possible error and no error conditions.

Gorman's fourth experiment was designed to more closely simulate the demand characteristics of the Eleusis task by having subjects again solve for an alternating rule, but the rule required that the sequences had to

alternate even-odd-even within themselves and across the series of sequences. As in the third experiment, subjects were given ten sequences and the test results and allowed to make five additional tests. A significantly greater number of subjects in the no error condition solved the task as compared to those in the possible error condition, though no significant differences between conditions were found for the number of expected incorrect trials and the number of repeated trials. While the analysis of the results is still incomplete, Gorman has speculated that in the first two experiments (when tests were independent and the opportunity to repeat trials was unlimited) subjects may have counteracted the effect of possible error on problem solution by increasing the number of trials and repeating disconfirming tests.

The psychological effects of a subject's knowledge of the possibility of system failure error should be distinguished from the effects of actual data error. Markowitz and Mynatt (1982) demonstrated that the presence of actual error increased the overall number of trials and the number of repeated trials, while decreasing successful solution rates. Gorman (1987) analyzed the effects of subjects' knowledge of possible error, but did not find a significant difference in solution rates between possible error and no error groups. It should be noted that unlike Markowitz and Mynatt, whose subjects were told that error would occur occasionally and were given actual system failure error, Gorman gave subjects a range of possible error (0-20%) but no actual error. Thus, subjects' information about the presence of error differed between the studies and that difference may have affected the results.

B. RESEARCH CONDUCTED UNDER THIS CONTRACT

Two experiments were completed for this phase of the contract. Experiment 7 was designed to compare problem-solving styles in both Single ("Dax/Not Dax") and Two-Rule ("Dax/Med") computerized versions of the Wason 2-4-6 rule induction task using three data feedback situations: (1) No Error (2) Informed Error and (3) Uninformed Error. The Two-Rule version ("Dax/"Med") was originally presented as part of a four-experiment study by Tweney et al. (1980), which attempted, through various instructional manipulations, to modify subjects' use of a positive test strategy and increase subsequent solving efficiency. The Two-Rule manipulation, following a method used earlier by Wetherick (1962), involved substituting the titles of "Dax" and "Med" for Wason's traditional "Right" or "Wrong" experimenter test-response categories. For example, if the rule was "three ascending numbers", the sequence "2, 4, 6" was responded to as a "Dax" and the sequence "6, 4, 2" as a "Med".

The structural change (solving for two rules) significantly increased solving efficiency, compared to Tweney et al.'s first three experiments which had used several variations in instructions from the Single-Rule task to increase disconfirmation. The Tweney et al. results suggested that finding two rules allowed the subjects to use a positive test strategy to solve the task more efficiently. In the Single-Rule task version, a potentially disconfirming datum was often ignored, apparently because it was considered a "wrong" answer, while in the Two-Rule version such a datum was considered relevant to the "Med" hypothesis. For example, a positive test (e.g. 1, 3, 5) of a plausible "Med" hypothesis (e.g. "odd numbers") would ultimately disconfirm the corresponding "Dax" hypothesis

(e.g. "even numbers") and expand the subject's knowledge base about "Dax" sequences. Thus, though a positive test strategy limited the subject's knowledge base in the Single-Rule task, in the Two-Rule task it expanded the range of possible evidence. Two later studies (Walker & Tweney, 1983; Walker, 1985) comparing Single and Two-Rule versions of the task under various instructional conditions have reported similar though somewhat less dramatic results.

Based in part on the results of Markowitz and Mynatt and of Gorman's recent findings, it was hypothesized that informing subjects of the presence of error in the Single-Rule task version should make subjects' hypotheses resistant to change by providing a rationale for ignoring disconfirming evidence. Conversely, subjects asked to find two rules should not ignore disconfirming evidence, even when informed of the presence of error, since a disconfirmatory trial could be incorrectly used to modify either or both rules.

Experiment 8 was designed to compare how informing or not informing subjects about the presence of error affects task performance under actual error and no error conditions. The 2 X 2 factorial design was based on the Experiment 7 methodology but used only the Single-Rule task version.

EXPERIMENT 7

Method

Subjects. Ninety Bowling Green State University students (34 freshman, 21 sophomores, 22 juniors, 7 seniors, and 6 graduate students; 56 females, 34 males) were recruited for the experiment. They were paid $3.50 each for their participation.

Procedure. The experiment was designed to compare the standard Wason (1960) 2-4-6 task, in which subjects are asked to find one number-sequencing rule, to the Tweney et al. (1980) version, in which subjects are asked to find two interrelated number-sequencing rules (DAX and MED), in each of three feedback conditions. The two main task conditions (Single and Two-Rule) were crossed with three test-response conditions: (1) No Error, (2) Informed Error--subjects received some erroneous feedback and were cautioned in the instructions that error might occasionally occur, and (3) Uninformed Error--subjects received some erroneous feedback, but were not cautioned that error might occur. Fifteen subjects were randomly assigned to each of the following six groups: (1) Single-Rule, No Error; (2) Two-Rule, No Error; (3) Single-Rule, Informed Error; (4) Two-Rule, Informed Error; (5) Single-Rule, Uninformed Error; and (6) Two-Rule, Uninformed Error.

Before starting an experimental session, the experimenter loaded one of six randomly-selected programs, which corresponded to the six experimental conditions, into each of four Apple Macintosh computers. All programs were designed to compare a three-digit keyboard entry (e.g., 1, 3, 5) to a general number-sequencing rule, three ascending numbers. For the four Error conditions (Single and Two-Rule, Informed and Uninformed), the programs also included a subroutine which was randomly activated for approximately 20% of the data entries. The subroutine reversed the computer response to a data entry so that a sequence that actually fit the "ascending numbers" rule was responded to as not fitting and vice-versa. For the Single and Two-Rule Informed Error conditions, an abbreviated list of instructions displayed on the computer monitor contained a warning to the subject that not all the computer responses to the number-sequence

tests were correct. For the single and two-rule uninformed error conditions, the abbreviated lists of instructions were the same as those used for the corresponding No Error conditions.

Upon entering the laboratory, subjects were asked to take a seat at one of the four computers. They were instructed to read a sheet of instructions concerning the task procedures, which corresponded to the One or Two-Rule task, depending on which task had been loaded into the computer (pp. 90 and 91). When all subjects had completed reading the instructions, the experimenter answered any questions that arose. Subjects were asked to write their name, age, class (freshman, sophomore, junior, senior, or graduate student), and major field of study at the top of the first record sheet provided next to the computer (see pp. 92 and 93). Each subject then entered his or her three initials from the keyboard as instructed by the display on the monitor screen and pressed the "Return" key. For the Single-Rule conditions, the program responded by displaying an abbreviated version of the task instructions, which included the sequence "2, 4, 6" as an example that fit the rule (see p. 94); by giving a highlighted warning about incorrect computer responses for the Informed Error condition (see p. 95); and by displaying a prompt "<?>" for the first data entry. Before subjects began testing sequences, the experimenter reminded them to enter the example, "2, 4, 6", on the top line of their response sheets under the "Number-Sequence Test" category and check the box under the "Fits" category. Subjects were then asked to write down any ideas they might have about the rule in the second space provided under the "Ideas About The Rule" category of the response sheet and to begin testing number-sequences.

Figure 2-1.

Single-Rule Instruction Sheet

WELCOME TO OUR LABORATORY

Today we would like you to play a game with the computer. The computer has been programmed to produce an infinite number of three-digit sequences. The object of the game is for you to find out what kind of number-sequencing rule the computer has been programmed to use. To discover the computer's rule, you may enter your own three-digit sequences from the keyboard. The computer will tell you if your sequences fit the rule or not. Based on the outcomes of these tests you should be able to find the rule.

On the table next to the computer you will find a record sheet which we would like you to use to keep track of your ideas, number-sequence tests, and the computer's responses. Following is a list of the steps you should follow in playing the game.

1. Type in your initials and press the <RETURN> key.

2. Read the instructions.

3. Enter any ideas you may have about what the rule might be and a three-digit sequence you would like to test on the record sheet.

4. Type in the three-digit sequence separating the digits with commas and press the <RETURN> key.

5. The computer will ask you if you think your three-digit sequence fits the actual rule. Type in "Y" for "YES", "N" for "NO", or "U" for "Unsure" or "I don't know" and press the <RETURN> key.

6. The computer will respond to your number-sequence test as either fitting or not fitting the rule. Record the computer's response on your record sheet.

7. Repeat Steps 3-6 until you are very sure you know what the rule is. When you think you know what the rule is, write it across the record sheet with the red pen and quietly raise your hand. The experimenter will tell you if your rule guess matches the actual rule.

8. If your guess was wrong, you may continue repeating Steps 3-6 and make another guess. You may repeat this process as many times as you wish.

If you have any questions while you are playing the game, just raise your hand and the experimenter will be glad to help you.

Figure 2-2.

Two-Rule Instruction Sheet

WELCOME TO OUR LABORATORY

Today we would like you to play a game with the computer. The computer has been programmed to produce two different lists of three-digit sequences. The object of the game is for you to find out what the two number-sequencing rules the computer has been programmed to use. To discover the computer's rules, you may enter your own three-digit sequences from the keyboard. The computer will tell you if your sequences fit the DAX or MED rules. Based on the outcomes of these tests you should be able to find the rules.

On the table next to the computer you will find a record sheet which we would like you to use to keep track of your ideas, number-sequence tests, and the computer's responses. Following is a list of the steps you should follow in playing the game.

1. Type in your initials and press the <RETURN> key.

2. Read the instructions.

3. Enter any ideas you may have about what the rules might be and a three-digit sequence you would like to test on the record sheet.

4. Type in the three-digit sequence separating the digits with commas and press the <RETURN> key.

5. The computer will ask you if you think your three-digit sequence fits the DAX rule. Type in "D" for "DAX", "M" for "MED", or "U" for "Unsure" or "I don't know" and press the <RETURN> key.

6. The computer will respond to your number-sequence test as either fitting the DAX or MED rule. Record the computer's response on your record sheet.

7. Repeat Steps 3-6 until you are very sure you know what the rules are. When you think you know what the rules are, write them across the record sheet with the red pen and quietly raise your hand. The experimenter will tell you if your rule guesses match the actual rules.

8. If your guesses were wrong, you may continue repeating Steps 3-6 and make another pair of guesses. You may repeat this process as many times as you wish.

If you have any questions while you are playing the game, just raise your hand and the experimenter will be glad to help you.

Figure 2-3.

Single-Rule Sample Response Sheet

IDEA ABOUT THE RULE	NUMBER-SEQUENCE TEST	FITS	DOESN'T FIT

Figure 2-4.

Two-Rule Sample Response Sheet

IDEAS ABOUT THE RULES	NUMBER-SEQUENCE TEST	DAX	MED

Figure 2-5.

Single-Rule Abbreviated Task Instructions (Monitor Display)

No Error and Uninformed Error Conditions

HI, I'M MAC, THE COMPUTER!
Would you please enter your initials and press <RETURN>
(subject's initials)
I have been programmed to generate an infinite list of three-digit sequences. I use a very general number-sequencing rule to get the job done. The object of the game we are going to play is for you to discover the number-sequencing rule I am using. I cannot tell you the rule, but I can tell you if a three-digit sequence that you enter from the keyboard fits my rule or not. For instance, if you were to give me the sequence--2,4,6--I would tell you that it fits my rule. You may test other three-digit sequences by entering three numbers separated by commas each time you see <?> on the screen. You may conduct as many number-sequence tests as you want. To make it easier for you to keep track of the number-sequences you have tried and my responses you should record them on the sheet next to the keyboard. When you are very sure you know what the rule is, just stop and write it across your test record sheet with the red pen and raise your hand. The experimenter will tell you if your guess is right. If your guess is wrong, you may continue to test more number sequences.
If you have any questions, please ask the experimenter now.

Figure 2-6.

Single-Rule Abbreviated Task Instructions (Monitor Display)

Informed Error Condition

HI, I'M MAC, THE COMPUTER!
Would you please enter your initials and press <RETURN>
(subject's initials)
I have been programmed to generate an infinite list of three-digit sequences. I use a very general number-sequencing rule to get the job done. The object of the game we are going to play is for you to discover the number-sequencing rule I am using. I cannot tell you the rule, but I can tell you if a three-digit sequence that you enter from the keyboard fits my rule or not. For instance, if you were to give me the sequence--2,4,6--I would tell you that it fits my rule. You may test other three-digit sequences by entering three numbers separated by commas each time you see <?> on the screen. You may conduct as many number-sequence tests as you want. To make it easier for you to keep track of the number-sequences you have tried and my responses you should record them on the sheet next to the keyboard. When you are very sure you know what the rule is, just stop and write it across your test record sheet with the red pen and raise your hand. The experimenter will tell you if your guess is right. If your guess is wrong, you may continue to test more number sequences. One word of CAUTION, once in a while I get mixed up and I may tell you a sequence fits when it doesn't and vice-versa.
If you have any questions, please ask the experimenter now.

After each three-digit keyboard entry, the Single-Rule programs responded by asking the subject to indicate whether he or she thought the test would fit the rule. For instance, if a subject's first hypothesis was "Even numbers" and the sequence "8, 10, 12" was entered, the screen displayed the question: "Do you think 8, 10, 12 will fit my rule?" The subject indicated his or her expectation of the computer response to the test by entering "Y" (yes), "No" (no), or "U" (unsure) and pressing <RETURN>. The entry of one of the three letters prompted the program to respond to the number-sequence test in one of two ways: (1) "That sequence fits my rule" or (2) "That sequence does not fit my rule".

For the Two-Rule conditions, the program responded by displaying an abbreviated version of the task instructions, which included the sequence "2, 4, 6" as an example that fit the "Dax" rule (see p. 97); by giving a highlighted warning about incorrect computer responses for the Informed Error condition (see p. 98); and by displaying a prompt "<?>" for the first data entry. Before subjects began testing sequences, the experimenter reminded them to enter the example, "2, 4, 6", on the top line of their response sheets under the "Number-Sequence Test" category and check the box under the "DAX" category. Subjects were then asked to write down any ideas they might have about the rules in the second space provided under the "Ideas About The Rules" category of the response sheet and to begin testing number-sequences.

Figure 2-7.

Two-Rule Abbreviated Task Instructions (Monitor Display)

No Error and Uninformed Error Conditions

HI, I'M MAC, THE COMPUTER!
Would you please enter your initials and press <RETURN>
(subject's initials)
I have been programmed to generate two different lists of three-digit sequences. I use two very general number-sequencing rules, DAX and MED, to get the job done. The object of the game we are going to play is for you to discover the two number-sequencing rules I am using. I cannot tell you the rules, but I can tell you if a three-digit sequence that you enter from the keyboard fits my DAX or MED rule. For instance, if you were to give me the sequence--2,4,6--I would tell you that it fits my DAX rule. You may test other three-digit sequences by entering three numbers separated by commas each time you see <?> on the screen. You may conduct as many number-sequence tests as you want. To make it easier for you to keep track of the number-sequences you have tried and my responses you should record them on the sheet next to the keyboard. When you are very sure you know what the rules are, just stop and write them across your test record sheet with the red pen and raise your hand. The experimenter will tell you if your guess is right. If your guess is wrong, you may continue to test more number sequences.
If you have any questions, please ask the experimenter now.

Figure 2-8.

Two-Rule Abbreviated Task Instructions (Monitor Display)

Informed Error Condition

HI, I'M MAC, THE COMPUTER!
Would you please enter your initials and press <RETURN>
(subject's initials)
I have been programmed to generate two different lists of three-digit sequences. I use two very general number-sequencing rules, DAX and MED, to get the job done. The object of the game we are going to play is for you to discover the two number-sequencing rules I am using. I cannot tell you the rules, but I can tell you if a three-digit sequence that you enter from the keyboard fits my DAX or MED rule. For instance, if you were to give me the sequence--2,4,6--I would tell you that it fits my DAX rule. You may test other three-digit sequences by entering three numbers separated by commas each time you see <?> on the screen. You may conduct as many number-sequence tests as you want. To make it easier for you to keep track of the number-sequences you have tried and my responses you should record them on the sheet next to the keyboard. When you are very sure you know what the rules are, just stop and write them across your test record sheet with the red pen and raise your hand. The experimenter will tell you if your guess is right. If your guess is wrong, you may continue to test more number sequences. One word of CAUTION, once in a while I get mixed up and may tell you a sequence fits my DAX rule when it doesn't and vice-versa.
If you have any questions, please ask the experimenter now.

After each three-digit keyboard entry, the Two-Rule programs responded by asking the subject to indicate which rule, DAX or MED, he or she thought the test fit. For instance, if a subject's first hypothesis about the "DAX" rule was "Even numbers" and the sequence "8, 10, 12" was entered, the screen displayed the question: "Do you think 8, 10, 12 will fit my DAX or MED rule?" The subject indicated his or her expectation of the computer response to the test by entering "D" (Dax), "M" (Med), or "U" (unsure) and pressing <RETURN>, which prompted the program to respond to the number-sequence test in one of two ways: (1) "That sequence fits my DAX rule" or (2) "That sequence fits my MED rule".

For all conditions, when a subject was ready to announce a rule or rules, as directed by the corresponding set of instructions, he or she wrote the rule or rules across the record sheet with a red pen and raised his or her hand. The experimenter, in turn, went to the subject's work station and indicated whether or not the announcement was correct by writing "Yes" or "No" next to the red entry. If the announcement was wrong, the subject was allowed to continue testing number-sequences and making announcements. Subjects were given 25 minutes to complete the task.

Results

Solution Rates

Solvers. The number of solvers was markedly less in the Informed and Uninformed Error conditions, than in the No Error conditions, in both the Single and Two-Rule task versions (see Table 2-1). Eleven of 60 subjects (18.3%) across all Error conditions eventually solved the task, compared to 27 of 30 subjects (90.0%) in both No Error conditions. For the Single-Rule task, 14 of 15 subjects (93.3%) in the No Error condition solved the

problem, compared to 8 of 30 (26.7%) in both Error conditions. The difference in the number solving the task on their first rule announcement, solving eventually but not on the first announcement, or not solving at all among the three Single-Rule conditions was significant ($\chi^2(4, N = 45) = 18.99, p<.001$). For the Two-Rule task, 13 of 15 subjects (86.7%) in the No Error condition solved the problem, compared to 3 of 30 subjects (10.0%) in both Error conditions. The difference in the number solving on their first rule announcement, solving eventually but not on the first announcement, or not solving at all among the three Two-Rule conditions was also significant ($\chi^2(4, N = 45) = 26.17, p<.001$). There was no significant difference in solving rates between the Single and Two-Rule, No Error conditions((2, N = 30) = 2.81, ns).

First-Announcement Solvers. The number solving the task in only one announcement was also much lower for both the Informed and Uninformed Error conditions across both the Single and Two-Rule task versions, compared to those in both No Error conditions (see Table 1).

Collapsing across Error conditions, 7 of 60 subjects (11.7%) solved the problem in one rule announcement, compared to 19 of 30 No Error subjects (63.3%). For the Single-Rule task, 4 of 30 Error subjects (13.3%) solved the problem in one rule announcement, compared to 8 of 15 No Error subjects (53.3%). The difference in the number of first-announcement solvers compared to eventual and non-solvers among the three Single-Rule conditions was significant ($\chi^2(2, N = 45) = 8.864, p<.025$). For the Two-Rule task, 3 of 30 Error subjects (10.0%) solved the task in one rule announcement, compared to 11 of 15 No Error subjects (73.3%). The difference in the number of first-announcement solvers compared to

eventual and non-solvers among the three Two-Rule conditions was also significant ($\chi^2(2, N = 45) = 18.871, p<.001$).

Table 2-1

Frequencies and Percentages of First-Announcement Solvers, Eventual Solvers and Non-solvers

Single-Rule Task Version (N = 45)

	1st-Announcement Solvers	Eventual Solvers	Non-Solvers
Condition	n (%)	n (%)	n (%)
No Error	8 (53.3%)	6 (40.0%)	1 (6.7%)
Informed Error	1 (6.7%)	3 (20.0%)	11 (73.3%)
Uninformed Error	3 (20.0%)	1 (6.7%)	11 (73.3%)
Totals	12 (26.7%)	10 (22.2%)	23 (51.1%)

Two-Rule Task Version (N = 45)

No Error	11 (73.3%)	2 (13.3%)	2 (13.3%)
Informed Error	2 (13.3%)	0 (00.0%)	13 (86.7%)
Uninformed Error	1 (6.7%)	0 (00.0%)	14 (93.3%)
Totals	14 (31.1%)	2 (4.4%)	29 (64.4%)j

Number-Sequence Tests: Frequencies. A two-way analysis of variance (task version x condition) indicated a significant difference in the mean

numbers of number-sequence tests subjects conducted among the No Error, Informed Error, and Uninformed Error conditions ($F(2, 84) = 12.425$, $p<.001$), but not between Single and Two-Rule task versions ($F(1, 84) = 2.527$, NS). No significant interaction between task version and condition was found. (See Table 2-2.) Subjects in the Error conditions conducted approximately twice as many number-sequence tests as No Error subjects. Across the four Error conditions, Error subjects conducted an average of 30.23 tests, while No Error subjects conducted an average of 16.15 tests. In the Single-Rule Informed and Uninformed Error conditions, subjects conducted an average of 29 tests, while Single-Rule No Error subjects conducted an average of 12.2 tests. In the Two-Rule Informed and Uninformed Error conditions, subjects conducted an average of 33.3 tests, while Two-Rule No Error subjects conducted an average of 20.1 tests.

Table 2-2

Mean Numbers of Tests Conducted.

	Task Version	
	Single-Rule	Two-Rule
Condition	Mean #	Mean #
No Error	12.2	20.1
Informed Error	28.7	34.5
Uninformed Error	29.3	28.4
Column Mean	23.4	27.7

Test Result Expectations. Comparisons of subjects' expected test result responses revealed significant differences between the Single and Two-Rule task versions for the percentages of total trials of "Yes" or "Dax" responses and "No" or "Med" responses. (See Table 2-3).

Table 2-3

Subjects' Expected Test Results.

Single-Rule Task Version

	Expect "Yes"	Expect "No"	"Unsure"
Condition	Mean # (%)	Mean # (%)	Mean # (%)
No Error	7.73 (63.4)	2.47 (20.2)	2.00 (16.4)
Informed Err	18.47 (64.4)	6.27 (21.9)	3.93 (13.7)
Uninformed Err	16.20 (55.2)	6.20 (21.1)	6.93 (23.6)
Column Means	14.13 (61.0)	4.98 (21.1)	4.29 (17.9)

Two-Rule Task Version

	Expect "Dax"	Expect "Med"	"Unsure"
No Error	7.27 (36.1)	9.00 (44.7)	3.87 (19.2)
Informed Err	11.93 (34.6)	11.80 (34.2)	10.80 (31.2)
Uninformed Err	11.93 (42.0)	13.00 (45.8)	3.47 (12.2)
Column Means	10.38 (37.6)	11.27 (41.5)	6.05 (20.9)

Note--Percentages computed as the percent of total trials.

A two-way analysis of variance (task version x condition) indicated

A two-way analysis of variance (task version x condition) indicated that the percentage of "Yes" or "Dax" responses was significantly different between the Single and Two-Rule task versions ($F(1, 84) = 28.832$, $p<.001$), but not among the No Error, Informed Error, and Uninformed Error conditions ($F(2, 84) = 1.141$, NS). No significant interaction effect was found. For the Single-Rule task version collapsing across conditions, subjects expected an average of 61.0% of the sequences tested to fit the experimenter's rule, while subjects in the Two-Rule version expected an average of 37.6% of the sequences to fit the analogous "Dax" rule.

The percentage of "No" or "Med" responses was also significantly different between the Single and Two-Rule task versions ($F(1, 84) = 41.486$, $p<.001$), but not among the No Error, Informed Error, and Uninformed Error conditions ($F(2, 84) = 1.089$, NS). For the Single-Rule task version across all conditions, subjects expected an average of 21.1% of the sequences not to fit the experimenter's rule, while subjects in the Two-Rule version expected an average of 41.5% of the sequences tested to fit the "Med" rule.

The percentage of "Unsure" responses was not significantly different between task versions, ($F(1, 84) = .113$, NS) or among the No Error, Informed Error, and Uninformed Error conditions ($F(2, 84) = .783$, NS). Across task versions, 17.9% of the Uninformed Error subjects' responses were "Unsure", compared to 17.8% of the No Error and 22.5% of the Informed Error subjects' responses.

Confirmation and Disconfirmation

Trial Outcome Categorization. For all conditions the outcome of each non-error trial was categorized as confirmatory, disconfirmatory, or

unclassifiable by comparing the subject's expected test result response ("Yes" or "Dax", "No" or "Med", or "Unsure") to the computer response to the test. A trial outcome was categorized as confirmatory if the expected test result response matched the computer response, disconfirmatory if the expected test result response did not match the computer response, and unclassifiable if the expected test result response was "Unsure" (see Table 2-4).

For the four Error conditions, the type of error trial (false positive or false negative in relation to the "ascending" rule) was also compared to the subject's expected test result response ("Yes" or "Dax", "No" or "Med", or "Unsure") to differentiate between spuriously confirmatory and spuriously disconfirmatory trial outcomes. An error trial outcome was categorized as spuriously confirmatory when the expected test result response matched the erroneous feedback or as spuriously disconfirmatory when the expected test result response did not match the erroneous feedback (see Table 2-4).

Comparisons. The percentages of total trial outcomes categorized as confirmatory (including spuriously confirmatory trials) were compared between the Single and Two-Rule task versions and across the No error, Informed Error, and Uninformed Error conditions. (See Table 2-5.) A two-way analysis of variance (task version x condition) indicated significant differences in the percentages of confirmatory trials between task versions ($F(1, 84) = 9.950$, $p < .01$) and among the three conditions ($F(2, 84) = 6.845$, $p < .01$). Of the total trial outcomes, 48.8% were confirmatory for the Single-Rule subjects, compared to 44.9% for the Two-Rule subjects. For No Error subjects, 56.5% of the total trial outcomes were confirmatory, compared to 40.1% for Informed Error

Criteria for Classification of Trial Outcomes

Single-Rule Task Version

Trial Classification	Test	Expect	Result
Confirmatory	1, 2, 3	"Yes"	"Yes"
Confirmatory	2, 2, 2	"No"	"No"
Disconfirmatory	1, 2, 3	"No"	"Yes"
Disconfirmatory	2, 2, 2	"Yes"	"No"
Spuriously Confirmatory	2, 2, 2	"Yes"	"Yes"
Spuriously Confirmatory	1, 2, 3	"No"	"No"
Spuriously Disconfirmatory	2, 2, 2	"No"	"Yes"
Spuriously Disconfirmatory	1, 2, 3	"Yes"	"No"

Two-Rule Task Version

Confirmatory	1, 2, 3	"Dax"	"Dax"
Confirmatory	2, 2, 2	"Med"	"Med"
Disconfirmatory	1, 2, 3	"Med"	"Dax"
Disconfirmatory	2, 2, 2	"Dax"	"Med"
Spuriously Confirmatory	2, 2, 2	"Dax"	"Dax"
Spuriously Confirmatory	1, 2, 3	"Med"	"Med"
Spuriously Disconfirmatory	2, 2, 2	"Med"	"Dax"
Spuriously Disconfirmatory	1, 2, 3	"Dax"	"Med"

subjects and 48.4% for Uninformed Error subjects. Post hoc pairwise comparisons of the differences between the means of the three conditions using Tukey's HSD test ($p<.01$) indicated the significant difference in the percentage of confirmatory trial outcomes was primarily between the No Error and Informed Error conditions.

Similarly, the percentages of disconfirmatory trial outcomes (including spuriously disconfirmatory trial outcomes) were compared using a two-way analysis of variance (task version x condition). Significant differences in the percentages of disconfirmatory trial outcomes were indicated between task versions ($F(1, 84) = 5.416$, $p<.025$) and among the No Error, Informed Error, and Uninformed Error conditions ($F(2, 84) = 15.154$, $p<.001$). Of the total trial outcomes, 32.9% were disconfirmatory for the Single-Rule subjects, compared to 33.3% for the Two-Rule subjects. For No Error subjects, 25.4% of the total trial outcomes were disconfirmatory, compared to 36.6% for Informed Error subjects and 33.6% for Uninformed Error subjects. Post hoc pairwise comparisons of the differences between the means of the three conditions using Tukey's HSD test ($p<.01$) indicated significant differences in the percentage of disconfirmatory trial outcomes between the No Error and Informed Error, as well as between the No Error and Uninformed Error conditions. A significant interaction effect (task version x condition) ($F(2, 84) = 8.463$, $p<.001$) for the percentages of disconfirmatory trial outcomes was also found. As shown in Table 2-5, the highest percentage of disconfirmatory trial outcomes (41.6%) for the Single-Rule subjects occurred in the Informed Error condition, while the highest percentage of disconfirmatory trial outcomes (38.5%) for the Two-Rule subjects occurred in the Uninformed Error condition.

Spurious Confirmation and Disconfirmation. In the Single-Rule, Informed Error condition, 32.0% of all error trial outcomes were categorized as spuriously confirmatory and 60.2% as spuriously disconfirmatory. In the Single-Rule, Uninformed Error condition, 21.9% of all error trial outcomes were categorized as spuriously confirmatory and

54.3% as spuriously disconfirmatory.

In the Two-Rule, Informed Error condition, 28.6% of all error trial outcomes were categorized as spuriously confirmatory and 34.5% as spuriously disconfirmatory. In the Two-Rule, Uninformed Error condition, 42.4% of all error trial outcomes were categorized as spuriously confirmatory and 48.5% as spuriously disconfirmatory.

Table 2-5

Percentages of Total Tests of Confirmation and Disconfirmation (%'s of Spurious Trials in Parentheses)

Single-Rule Task Version

	Confirmation		Disconfirmation	
Condition	Total	(Spurious)	Total	(Spurious)
No Error	61.7%	NA	21.9%	NA
Informed Err	44.6%	(7.7%)	41.6%	(14.4%)
Uninformed Err	47.5%	(5.2%)	28.9%	(12.6%)
Two-Rule Task Version				
No Error	53.3%	NA	27.5%	NA
Informed Err	36.3%	(6.6%)	32.4%	(7.9%)
Uninformed Err	49.3%	(9.9%)	38.5%	(11.3%)

Test Repetitions

Within each individual test protocol, a test was classified as a repetition if it duplicated a previous test. The number of subjects

repeating number-sequence tests was higher in the Error conditions, in contrast to the number of subjects in the No Error conditions (see Table 2-6). Across the Single and Two-Rule task versions, 41 of 60 Error subjects (68.3%) repeated tests, compared to 11 of 30 No Error subjects (36.7%). In the Single-Rule task version, 22 of 30 Error subjects (73.3%) repeated tests, compared to 3 of 15 No Error subjects (20.0%). The difference in the number of subjects repeating tests among the Single-Rule No Error, Informed Error, and Uninformed Error conditions was significant ((2, N = 45) = 11.52, $p<.01$). In the Two-Rule task version, 19 of 30 Error subjects (63.3%) repeated number-sequence tests, compared to 8 of 15 No Error subjects (53.3%). However, the difference in the number of subjects repeating tests among the Two-Rule No Error, Informed Error, and Uninformed Error conditions was not significant ((2, N = 45) = 0.556, NS).

Table 2-6

Number of Subjects Repeating Tests (%'s of Subjects/Condition in Parentheses)

	Task Version	
	Single-Rule (N = 45)	Two-Rule (N = 45)
Condition	n (%)	n (%)
No Error	3 (20.0%)	8 (53.3%)
Informed Error	11 (73.3%)	10 (66.7%)
Uninformed Error	11 (73.3%)	9 (60.0%)

Among all Single-Rule conditions, the mean number of repeated tests was 5.2 (12.3 percent of the total tests conducted). (See Table 2-7). Since the variances for the three Single-Rule samples, as well as the Two-Rule samples, were extremely unequal, thus violating the assumption of homogeneity of variance for basic ANOVA, the differences were analyzed using the Kruskal-Wallis H test. The difference in the mean numbers of repeated tests among the Single-Rule conditions was not significant ($H(2, N = 25) = 1.786$, $p = .409$, NS). Among all Two-Rule conditions, the mean number of repeated tests was 9.5 (20.9 percent of the total tests conducted). The difference in the mean numbers of repeated tests among the Two-Rule conditions was also not significant ($H(2, N = 27) = 4.676$, $p = .097$, NS).

Table 2-7

Mean Numbers of Repeated Tests (%'s of Total Tests in Parentheses)

	Task Version	
	Single-Rule	Two-Rule
Condition	Mean * (%)	Mean * (%)
No Error	2.3 (3.8%)	2.0 (5.3%)
Informed Error	5.4 (13.7%)	16.3 (32.8%)
Uninformed Error	5.5 (13.9%)	8.6 (18.1%)
Mean Totals	5.1 (12.1%)	9.5 (20.9%)

The type of repeated test (confirmatory or disconfirmatory) was based on the outcome of the trial being repeated. Across all Single-Rule

conditions, 17 of 25 subjects (68.0%) repeating tests repeated both confirmatory (34.6% of all repetitions) and disconfirmatory (42.5% of all repetitions) outcomes. Across all Two-Rule conditions, 16 of 27 subjects (59.3%) repeating trials repeated both confirmatory (31.3% of all repetitions) and disconfirmatory (36.4% of all repetitions) outcomes (See Table 2-8).

Table 2-8

Number of Subjects Repeating Confirmatory, Disconfirmatory, or Both Types of Tests (%'s of Subjects/Condition in Parentheses)

Single-Rule Task Version (N = 45)

	Types of Repeated Tests		
	Confirmatory	Disconfirmatory	Both
Condition	n (%)	n (%)	n (%)
No Error	0 (0.0%)	0 (0.0%)	3 (100.0%)
Informed Error	1 (9.1%)	3 (27.2%)	7 (63.7%)
Uninformed Error	1 (9.1%)	3 (27.2%)	7 (63.7%)
Column Totals	2 (8.0%)	6 (24.0%)	17 (68.0%)
Two-Rule Task Version (N = 45)			
No Error	6 (75.0%)	1 (12.5%)	1 (12.5%)
Informed Error	2 (20.0%)	1 (10.0%)	7 (70.0%)
Uninformed Err	1 (11.1%)	0 (00.0%)	8 (88.9%)
Column Totals	9 (33.3%)	2 (7.4%)	16 (59.3%)

Discussion

Compared to the No Error conditions for both the Single and Two-Rule task versions, the significantly higher number of tests and decreased solving rates for the Informed and Uninformed Error conditions indicated that data error seriously disrupted task performance. The findings support Markowitz and Mynatt's (1982) results in which it was reported that feedback error disrupted solving efficiency. It should also be noted that the percentage (73.3%) of first-time solvers for the Two-Rule, No Error condition was higher than Tweney et al.'s (1980) findings, in which 56.8% of the Two-Rule subjects solved on their first announcement. Similarly, the percentage (53.3%) of first-time solvers for the Single-Rule, No Error condition was higher than Wason's (1966) original findings, in which only 20.7% of the subjects solved on their first rule announcement. The difference in solution rates between the current and earlier studies might be partially attributable to automation of the task, which, by limiting the amount of experimenter/subject interaction, allowed a more efficient means of testing sequences and receiving results than the usual paper and pencil method of task administration.

The primary purpose of Experiment 7, however, was not to demonstrate whether or not error disrupted task performance, but how informing or not informing subjects about error would affect hypothesis-testing heuristics. According to Tweney and Doherty (1983), "error-free and error-prone data (should) elicit different hypothesis testing strategies" (p. 153), such as conducting more potentially confirmatory tests, ignoring disconfirmatory results, and selectively repeating more disconfirmatory tests. The results of the Kern (1982) and

Gorman (1986) studies using other tasks had also indicated that disconfirmatory results were frequently ignored and/or replicated when the possibility of error was introduced. The results of Experiment 7 demonstrated that Single and Two-Rule Error subjects conducted more tests overall and were more likely to repeat tests than No Error subjects. However, no significant differences were found in the percentages of subjects' expectations ("Yes", "No", "Unsure") regarding test results between the Single-Rule Informed and Uninformed Error conditions, as well as between the Single-Rule No Error and Error conditions (see Table 2-3). Subjects in all three Single-Rule conditions expected approximately three times as many tests to fit the experimenter's rule as not fit. Similarly, the percentages of Two-Rule subjects' expectations ("Dax", "Med", "Unsure") between Informed and Uninformed Error conditions, as well as between No Error and Error conditions were not significantly different. In contrast to the Single-Rule subjects, Two-Rule subjects expected similar percentages of tests to fit both "Dax" and "Med".

Though it was expected that Informed Error subjects would be more suspicious of the data and more likely to repeat tests than Uninformed Error subjects, Single-Rule subjects in both the Informed and Uninformed Error conditions were equally likely to repeat number-sequence tests. In each of the Single-Rule Error conditions, 73.3% of the subjects repeated tests, compared to only 20.0% of the No Error subjects. Interestingly, the percentage of subjects (53.3%) in the Two-Rule, No Error condition repeating tests was very similar to the percentage of repeaters (63.3%) in both Two-Rule Error conditions. Whether or not subjects repeated tests appeared to be more closely related to when disconfirmation occurred and to the proportionate balance between confirmation and disconfirmation.

Though both Single-Rule No Error and Error subjects expected similar proportions of tests to fit and not fit the experimenter's rule, the tests conducted by Error subjects resulted in a lower proportion of confirmatory to disconfirmatory trial outcomes than tests conducted by the No Error subjects. Single-Rule, No Error subjects obtained approximately three times as much confirmation as disconfirmation from their tests, while Informed Error subjects obtained an approximately equal proportion of confirmation and disconfirmation, and Uninformed Error subjects obtained less than twice as much confirmation as disconfirmation. Also, for most Single-Rule No Error subjects, the majority of confirmatory trial outcomes was obtained early in the testing process with disconfirmation coming later, after a hypothesis had been well-established. However, for 17 of 30 (56.7%) Single-Rule Error subjects, their first hypothesis was spuriously disconfirmed within the first five trials. Such early spurious disconfirmation, before a hypothesis could be well-confirmed, may have confused Single-Rule Informed and Uninformed Error subjects. As Klayman and Ha (1987) have suggested, confirmatory heuristics, such as a positive test strategy, are especially powerful in task situations in which the validity of feedback is questionable, but in the present task subjects who attempted to confirm or establish ideas by using a positive test strategy were quickly misled.

As noted earlier for the Two-Rule task the proportions of tests expected to fit and not fit the experimenter's rules were similar among the Two-Rule No Error and Error conditions with subjects' test result expectations divided almost equally between both rules. The difference between the two task versions in subjects' test expectations demonstrated how finding two rules, rather than one, differentially

affects the task structure. The proportions of confirmatory and disconfirmatory trial outcomes, unlike the Single-Rule conditions, were also very similar among the No Error and Error conditions, as were the percentages of subjects repeating tests. Thus Two-Rule subjects in all three conditions, as did Single-Rule Error subjects, received less confirmation in relation to disconfirmation and were more likely to repeat tests than Single-Rule No Error subjects.

The lack of differences in subjects' expectations about test results within each task version indicated that neither informing subjects about error nor actual error affected subjects' basic test strategies. However, the presence of error so severely disrupted task performance that the lack of differences between the Informed and Uninformed Error conditions may have been due to a "floor effect". Therefore, Experiment 2 was designed to differentiate between the psychological effects of informing subjects about error and actual error on hypothesis-testing heuristics using only the single rule condition.

EXPERIMENT 8

Method

Subjects. Eighty Bowling Green State University introductory psychology students (54 freshmen, 24 sophomores, and 2 juniors; 58 females, 22 males) were recruited for the study. They were paid $5.00 each for their participation.

Procedure. The experiment utilized a 2 x 2 factorial design in which system failure (Error) vs. no system failure (No Error) feedback conditions were crossed with error warning (Informed) vs. no error warning (Uninformed) instructions. Twenty subjects were randomly assigned to

each of the following separate groups: (1) No Error; (2) Informed No Error; (3) Informed Error and (4) Uninformed Error.

Before starting an experimental session, the experimenter loaded one of four randomly selected programs into each of four Apple Macintosh computers. The programs were duplicates of the programs used for the Experiment 7 Single-Rule No Error and Error conditions. The Experiment 7 laboratory procedure was followed, including the use of the Single-Rule Instruction and Response sheets (see pp. 90 and 92). For the two Informed conditions (Error and No Error), the warning to the subject that not all the computer responses to the number-sequence tests were correct was enhanced by capitalizing the entire sentence (see p.117). As in Experiment 7, for the two Uninformed conditions (Error and No Error), the abbreviated list of instructions did not contain an error warning (See p.94). All subjects were again given 25 minutes to complete the task. At the end of the experimental session, the experimenter gave the correct solution, explained the purpose of the experiment, paid the subjects, and distributed debriefing forms.

Results

Solution Rates

Solvers. As in Experiment 7, the number of solvers was considerably lower for both the Informed and Uninformed Error conditions, compared to the No Error and Informed No Error conditions (see Table 2-9). Only 5 of 40 subjects (12.5%) in both the Informed and Uninformed Error conditions were able to solve the task, compared to 35 of 40 No Error and Informed No Error subjects (87.5%). The differences in the number of subjects solving the task on their first rule announcement, eventually, or not at all among the four conditions were significant ($\chi^2(6, N = 80) = 56.633$, $p<.001$).

Figure 2-9.

Abbreviated Task Instructions (Monitor Display)

Informed No Error and Informed Error Conditions

HI, I'M MAC, THE COMPUTER!
Would you please enter your initials and press <RETURN>
(subject's initials)
I have been programmed to generate an infinite list of three-digit sequences. I use a very general number-sequencing rule to get the job done. The object of the game we are going to play is for you to discover the number-sequencing rule I am using. I cannot tell you the rule, but I can tell you if a three-digit sequence that you enter from the keyboard fits my rule or not. For instance, if you were to give me the sequence, 2,4,6, I would tell you that it fits my rule. You may test other three-digit sequences by entering three numbers separated by commas each time you see <?> on the screen. You may conduct as many number-sequence tests as you want. To make it easier for you to keep track of the number-sequences you have tried and my responses you should record them on the sheet next to the keyboard. When you are very sure you know what the rule is, just stop and write it across your test record sheet with the red pen and raise your hand. The experimenter will tell you if your guess is right. If your guess is wrong, you may continue to test more number sequences.
ONE WORD OF CAUTION, ONCE IN A WHILE I GET MIXED UP AND I MAY TELL YOU A SEQUENCE FITS WHEN IT DOESN'T AND VICE-VERSA.
If you have any questions, please ask the experimenter now.

Solution rates (first-announcement solvers, eventual solvers, and non-solvers) between Informed conditions, No Error and Error, also differed significantly ($\chi^2(2, N = 40) = 12.241$, $p<.01$). However, the differences in solution rates did not differ between the Informed and Uninformed Error conditions ($\chi^2(2, N = 40) = 5.714$, NS) or between the No Error and Informed No Error conditions ($\chi^2(2, N = 40) = 6.739$, NS).

First-Announcement Solvers. The number of first announcement solvers was much lower in the Informed and Uninformed Error conditions than in the No Error and Informed No Error conditions (see Table 2-9). Across Informed and Uninformed Error conditions, 3 of 40 subjects (7.5%) solved the problem in one rule announcement, compared to 24 of 40 No Error and Informed No Error subjects (60.0%). The difference in the number of first-announcement solvers compared to eventual and non-solvers among the four conditions was significant ($\chi^2(3, N = 80) = 32.816$, $p<.001$).

Table 2-9

Frequencies and Percentages of First-Announcement Solvers, Eventual Solvers and Non-solvers (N = 80)

	First-Announcement	Eventual	Non-solvers
Condition	n (%)	n (%)	n (%)
No Error	16 (80.0%)	3 (15.0%)	1 (5.0%)
Informed No Error	8 (40.0%)	8 (40.0%)	4 (20.0%)
Informed Error	3 (15.0%)	2 (10.0%)	15 (75.0%)
Uninformed Error	0 (00.0%)	0 (00.0%)	20 (100.0%)

Number-Sequence Tests

Frequencies. The mean numbers of number-sequence tests conducted by subjects differed significantly among the four conditions ($F(3, 76) = 19.89$, $p<.001$) (See Table 2-10). Subjects in the Error conditions conducted more tests than No Error and Informed No Error subjects. Informed and Uninformed Error subjects conducted an average of 28.95 tests, while No Error and Informed No Error subjects averaged 14.6 tests.

Table 2-10

Mean Numbers of Tests Conducted

Condition	Mean #
No Error	14.50
Informed No Error	14.70
Informed Error	28.85
Uninformed Error	29.05
Mean	21.78

Test Result Expectations. An ANOVA among the four conditions indicated no difference in the percentages of total trials of "Yes" responses to the "Expected Test Result" category ($F(3, 76) = .234$, NS). (See Table 2-11.) Across the No Error and Informed No Error conditions, subjects responded that they expected 51.25% of the total trials to fit the rule, compared to 58.7% of the total trials attempted by Informed and Uninformed Error subjects. The percentages of "No" responses among the four conditions

were also not significantly different (F (3, 76) = .757, NS). Across the No Error and Informed No Error conditions, subjects responded that they did not expect 20.2% of the total trials to fit the rule, compared to 17.2% of the total trials attempted by Informed and Uninformed Error subjects. Similarly, the percentages of "Unsure" responses among the conditions did not differ among the four conditions (F(3, 76) = .619, NS). Across the No Error and Informed No Error conditions, subjects responded that they were unsure about 27.0% of the test results, compared to 25.3% of the total trials attempted by Informed and Uninformed Error subjects.

Table 2-11

Subjects' Expected Test Results

	Expected "Yes"	Expected "No"	"Unsure"
Condition	Mean * (%)	Mean * (%)	Mean * (%)
No Error	7.90 (54.5%)	3.15 (21.7%)	3.45 (23.8%)
Informed No Error	7.05 (48.0%)	2.30 (15.6%)	5.35 (36.4%)
Informed Error	17.30 (60.0%)	5.65 (19.6%)	5.90 (20.4%)
Uninformed Error	16.65 (57.3%)	4.50 (15.5%)	7.90 (27.2%)
Mean Totals	12.23 (56.1%)	3.90 (17.9%)	5.65 (26.0%)

Confirmation and Disconfirmation

Trial Outcome Categorization. As in Experiment 7, for all conditions the outcome of each non-error trial was categorized as confirmatory, disconfirmatory, or unclassifiable by comparing the subject's expected test result response ("Yes", "No", or "Unsure"). A trial outcome was

categorized as confirmatory if the expected test result response matched the computer response, disconfirmatory if the expected test result response did not match the computer response, and unclassifiable if the expected test result response was "Unsure". (See Table 2-4, p. 106.) For the two Error conditions, the type of error trial (false positive or false negative in relation to the "ascending" rule) was also compared to the subject's expected test result response ("Yes", "No", or "Unsure") to differentiate between spuriously confirmatory and spuriously disconfirmatory trial outcomes. The outcome of an error trial was categorized as spuriously confirmatory when the expected test result response matched the erroneous feedback or as spuriously disconfirmatory when the expected test result response did not match the erroneous feedback (see Table 2-4).

The percentages of total test outcomes categorized as confirmatory (including spuriously confirmatory trial outcomes) were compared across the four conditions. (See Table 2-12.) An analysis of variance indicated no significant difference in the percentages of confirmatory trials among the No Error, Informed No Error, Informed Error, and Uninformed Error conditions ($F(3, 76) = 1.194$, NS). Of the total test outcomes, 54.3% were confirmatory for No Error and Informed No Error subjects and 47.0% were confirmatory for Informed and Uninformed Error subjects.

A second analysis of variance indicated a significant difference in the percentages of disconfirmatory trial outcomes among the four conditions ($F(3, 76) = 9.983$, $p<.001$). For No Error and Informed No Error subjects, 15.6% of the total test outcomes were disconfirmatory, compared to 29.2% for Informed and Uninformed Error subjects. Post hoc pairwise comparisons of the differences between the means of the four conditions

using Tukey's HSD test ($p<.01$) indicated significant differences in the percentage of disconfirmatory trial outcomes between the Informed No Error condition and the Informed and Uninformed Error conditions.

Table 2-12

Percentages of Total Tests of Confirmation and Disconfirmation (%'s of Spurious Trials in Parentheses)

	Confirmation	Disconfirmation
Condition	Total (Spurious)	Total (Spurious)
No Error	55.9% NA	20.3% NA
Informed No Error	52.7% NA	10.9% NA
Informed Error	49.1% (5.5%)	30.5% (14.6%)
Uninformed Error	44.8% (7.2%)	27.9% (12.6%)

Spurious Confirmation and Disconfirmation. In the Informed Error condition, 22.4% of all error trial outcomes were categorized as spuriously confirmatory, 58.7% as spuriously disconfirmatory. In the Uninformed Error condition, 29.0% of all error trial outcomes were spuriously confirmatory and 50.3% as spuriously disconfirmatory.

Test Repetitions. The difference in the number of subjects repeating tests among the No Error, Informed No Error, Informed Error and Uninformed Error conditions was significant ($\chi^2(3, N = 80) = 29.039$, $p<.005$). (See Table 2-13.) The number of subjects repeating tests was higher in the Informed and Uninformed Error Conditions than in the No Error and Informed No Error conditions. Across Informed and Uninformed Error

conditions, 28 of 40 subjects (70.0%) repeated trials, compared to 7 of 40 No Error and Informed No Error subjects (17.5%).

Table 2-13

Number of Subjects Repeating Tests (%'s of Subjects/Condition in Parentheses, N = 80)

	Repeating	Not Repeating
Condition	n (%)	n (%)
No Error	4 (20.0%)	16 (80.0%)
Informed No Error	3 (15.0%)	17 (85.0%)
Informed Error	16 (80.0%)	4 (20.0%)
Uninformed Error	12 (60.0%)	8 (40.0%)
Column Totals	35 (43.8%)	45 (56.2%)

Among the four conditions, the mean number of repeated tests was 2.6 (12.1% of the total tests conducted). (See Table 2-14.) As in Experiment 7, the variances for the samples were extremely unequal and the differences between the means were analyzed using the Kruskal-Wallis H test. The difference in the mean numbers of repeated tests among the four conditions was not significant ($H(3, N = 37) = 7.09$, $p=.069$, NS). The type of repeated test (confirmatory or disconfirmatory) was based on the outcome of the trial being repeated. Across all conditions, 17 of 35 subjects (48.6%) repeating tests, repeated both confirmatory (36.4% of all repeated tests) and disconfirmatory (43.5% of all repeated tests) test

outcomes, 8 (22.9%) repeated only confirmatory test outcomes, and 10 (28.6%) repeated only disconfirmatory test outcomes. (See Table 2-15.)

Table 2-14

Mean Numbers of Repeated Tests (%'s of Total Tests in Parentheses)

Condition	Mean * (%)
No Error	1.3 (1.7%)
Informed No Error	1.3 (1.4%)
Informed Error	5.8 (20.1%)
Uninformed Error	4.3 (14.8%)
Mean Totals	3.2 (9.5%)

Table 2-15

Number of Subjects Repeating Confirmatory, Disconfirmatory, or Both Types of Tests (%'s of Subjects Repeating in Parentheses, N = 80)

	Type of Repeated Tests		
	Confirmatory	Disconfirmatory	Both
Condition	n (%)	n (%)	n (%)
No Error	3 (75.0%)	1 (25.0%)	0 (0.0%)
Informed No Error	2 (66.7%)	1 (33.3%)	0 (0.0%)
Informed Error	3 (18.8%)	4 (25.0%)	9 (56.2%)
Uninformed Error	0 (0.0%)	4 (33.3%)	8 (66.7%)
Column Totals	8 (22.9%)	10 (28.6%)	17 (48.5%)

Discussion

Subjects in both Informed and Uninformed Error conditions, as in the Single-Rule task version of Experiment 7, used a significantly higher number of tests and had substantially decreased solving rates compared to subjects in the No Error and Informed No Error conditions. The results again indicated that data error seriously disrupts task performance. The percentage (80.0%) of first-time solvers for the No Error condition was again higher than Wason's (1960) original findings, as well as higher than the percentage (53.3%) of first-time solvers reported in Experiment 7. The total percentage (95.0%) of both first-time and eventual solvers for the No Error condition was very similar to the total percentage (93.4%) reported for the Single-Rule No Error condition in Experiment 7. As shown in Table 2-16, the Experiment 8 findings for the No Error, Informed Error, and Uninformed Error conditions not only replicated the Experiment 7 Single-Rule task version findings for solution rates, but also the mean number of trials, and the mean numbers and percentages of subjects' expected test result responses. Furthermore, the number of subjects repeating tests, as well as the mean number of repeated tests, as found in Experiment 7 were replicated in Experiment 8.

Table 2-16

Comparison of Experiment 7 and Experiment 8 Findings

Solution Rates

Condition	First Solvers n (%)	Eventual Solvers n (%)	Non-Solvers n (%)
No Error (7)	8 (53.3%)	6 (40.0%)	1 (6.7%)
No Error (8)	16 (80.0%)	3 (15.0%)	1 (5.0%)
Informed Err (7)	1 (6.7%)	3 (20.0%)	11 (73.3%)
Informed Err (8)	3 (15.0%)	2 (10.0%)	15 (75.0%)
Uninformed Err (7)	3 (20.0%)	1 (6.7%)	11 (73.3%)
Uninformed Err (8)	0 (00.0%)	0 (00.0%)	20 (100.0%)

Mean Numbers of Tests Conducted

Condition	Experiment 7 Mean No.	Experiment 8 Mean No.
No Error	12.2	14.5
Informed Error	28.7	28.9
Uninformed Error	29.3	29.1

Table 2-16 cont.

Subjects' Expected Test Results

Condition	Expect "Yes" Mean # (%)	Expect "No" Mean # (%)	"Unsure" Mean # (%)
No Error (7)	7.73 (63.4)	2.47 (20.2)	2.00 (16.4)
No Error (8)	7.90 (54.5)	3.15 (21.7)	3.45 (23.8)
Informed Err (7)	18.47 (64.4)	6.27 (21.9)	3.93 (13.7)
Informed Err (8)	17.30 (60.0)	5.65 (19.6)	5.90 (20.4)
Uninformed Err (7)	16.20 (55.2)	6.20 (21.1)	6.93 (23.6)
Uninformed Err (8)	16.65 (57.3)	4.50 (15.5)	7.90 (27.2)

The primary purpose of Experiment 8 was to differentiate between the psychological effects on hypothesis-testing heuristics of informing subjects of the possibility of error and the effects of actual error. As in Experiment 7, error did not increase subjects' use of a positive test strategy. For instance, across all conditions the percentages of subjects' expectations ("Yes", "No", "Unsure") regarding test results were very similar with subjects expecting almost three times as many tests to fit, as not fit, the rule. However, Error subjects were more likely to use a greater number of trials and to repeat tests than No Error and Informed No Error subjects. Thus error, but not informing subjects of the possibility that error might occur, affected hypothesis-testing heuristics by increasing overall testing and repetition.

Compared to No Error subjects, Informed No Error subjects used a

similar number of tests and were not likely to repeat tests, but less likely to solve the task on their first rule announcement. Only 8 of 20 (40.0%) Informed No Error subjects solved the task on their first announcement, compared to 16 of 20 (80.0%) No Error subjects. Several factors, such as uncertainty about the reliability of the data, a preference for initially using a positive test strategy, and the availability of reliable rule feedback from the experimenter, might have contributed to the lower first-announcement solution rates for the Informed No Error subjects. For instance, Informed No Error subjects had a higher percentage (36.4%) of "Unsure" test result expectations than No Error subjects (23.8%), indicating that the possibility of error increased subjects' uncertainty about test results. Of the eight Informed No Error eventual solvers, two (25.0%) responded as "Unsure" about the results for the majority of tests conducted, while three (37.5%) expected and received only confirmatory results and two (25.0%) received one disconfirmatory result before making their first rule announcement. The latter two subjects modified their original hypotheses to include the disconfirmatory results, but did not completely abandon their original ideas. Thus, it appeared that the majority of these eventual solvers based incorrect first announcements primarily on confirmatory test results and disconfirmed original hypotheses by announcing them as rules.

As in Experiment 7, Error subjects who attempted to establish an initial hypothesis using a positive test strategy were quickly misled by early spurious disconfirmation. In contrast, Informed No Error subjects, somewhat uncertain about data quality, were able to use a positive test strategy to develop well-confirmed hypotheses and then depend on the reliability of experimenter rule feedback to confirm or disconfirm them.

Walker and Tweney (1986) recently compared restricting or not restricting the number of rule announcements and demonstrated that subjects were less likely to use a negative test strategy and more likely to announce well-confirmed hypotheses as rules when no restrictions were placed on the number of rule announcements. Any further error research using this or other hypothesis-testing tasks should be designed to minimize the use any source of information other than data collection for hypothesis testing (e.g., permitting only one rule announcement).

As demonstrated in both experiments, subjects given error conducted a significantly greater number of tests and were more likely to repeat tests than subjects given no error, supporting Kern's (1982) and Markowitz and Mynatt's (1982) general findings. The greater number of trials and higher number of subjects repeating tests in the Error conditions also indicated that subjects were attempting to cope with the confusing combinations of sequences produced by both spuriously confirmatory and disconfirmatory trials. However, in contrast to the Kern and Markowitz and Mynatt studies, Error subjects did not demonstrate a preference for replicating only disconfirmatory trials. Instead, Error subjects appeared to become suspicious of the data overall and replicated both confirmatory and disconfirmatory trials. Increased repetition in the Error conditions, combined with the greater number of tests conducted, resulted in building a large and confusing data base contaminated by both spuriously confirmatory and disconfirmatory results.

The study also indicated that, in general, subjects did not increase their use of a positive test strategy when the possibility of data error was introduced. The findings of Experiment 8 support Gorman's recent results in which subjects who were informed about, but not given actual

error, did not spend a great deal of time replicating tests to check for error. However, unlike Gorman's subjects, Informed No Error subjects in the current study did not test significantly more sequences than No Error subjects to rule out the possibility of error. As noted earlier, the primary difference between the No Error and Informed No Error conditions was found in how subjects verified their initial hypotheses. Many Informed No Error subjects first established a well-confirmed hypothesis, appearing to ignore the possibility that some confirmatory trials might be erroneous, and made a premature incorrect rule announcement. Gorman's study, on the other hand, restricted rule announcements, which might have compelled subjects to conduct more tests to check for error.

Most successful subjects in both experiments given the Single-Rule task version used a combination of early confirmation followed by later disconfirmation to solve the task. This finding supported the results of Walker and Tweney (1983) and Walker (1985, 1986).

In general the study has shown that actual system failure error decreased the likelihood that a well-confirmed hypothesis could be established, in part at least, due to spuriously disconfirmatory results. In contrast to the No Error conditions, the introduction of system failure error, both as a possibility and as an actual occurrence, affected the use and utility of confirmatory heuristics in the process of hypothesis discovery. Thus, though a confirmatory heuristic was useful for establishing a hypothesis, it was an inefficient method for eventually solving the task unless combined with some attempts to disconfirm and systematic replication for determining the extent of error present in the data.

PART 3

RESEARCH ON DATA ERROR USING DATA SELECTION PARADIGMS

CONTENTS

A. PRIOR RESEARCH

In a recent review, Fischoff and Beyth-Marom (1983) concluded that the most powerful of several "metabiases" which are exhibited by people attempting to test hypotheses "is the tendency to ignore P(D/not-H) when evaluating evidence" (p.257). In a simple case in which this metabias operates, subjects are asked to choose between two hypotheses (H and not-H), based on two data, D_1 and D_2. They are given the values of P(H) and $P(D_1/H)$, and are then asked to select one more conditional probability from either $P(D_1/\text{not-H})$ or $P(D_2/H)$. On a Bayesian analysis, the normative solution to such a problem is to choose both $P(D_1/H)$ and $P(D_1/\text{not-H})$. This is because likelihood ratio, P(D/H)/P(D/not-H), necessary to calculate the posterior odds that H is true (that is, the likelihood that H rather than not-H is the case following receipt of the datum D), requires both values. Since the diagnosticity of a given datum D is determined by the extent to which the likelihood ratio deviates from 1.0, diagnosticity is unknown unless both $P(D_1/H)$ and $P(D_1/\text{not-H})$ are known. Thus, failure to select $P(D_1/\text{not-H})$ is non-normative; it is, however, very common.

The failure to select P(D/not-H) can be seen as a failure to consider alternative hypotheses. Failure to consider alternatives has been demonstrated across a number of tasks and subject populations. For example, a generalization in both the concept formation and problem solving literatures is that people tend to focus on one hypothesis at a time (e.g., Newell & Simon, 1972, p. 752). The same tendency is also vividly

shown in research on theory perseverance (e.g., Jennings, Amabile, & Ross, 1982) and illusory correlation (e.g., Smedslund, 1963). It has been shown in relatively simple tasks such as Wason's (1960) 2-4-6 problem and in more complex situations such as simulations of scientific problems (Mynatt, Doherty, & Tweney, 1978). Subjects as diverse as college students (Mynatt, Doherty, & Tweney, 1977), Protestant ministers (Mahoney & DeMonbreun, 1978), advanced medical students (Kern & Doherty, 1982), and scientists (Mitroff, 1974) have exhibited it.

The research most relevant to the present report is that involving the "pseudodiagnosticity" paradigm (Doherty, Mynatt, Tweney, & Schiavo, 1979; Doherty, Schiavo, Tweney, & Mynatt, 1981; Kern & Doherty, 1982; Beyth-Marom & Fischoff, 1983). In these studies subjects are typically given a choice between P(D_1/not-H) and information which is diagnostically worthless. For example, Doherty, et al., (1979) asked subjects to decide from which of two islands an archaeological find (a pot) had come. They were given six binary characteristics of the find (e.g., that it had a curved handle) and were allowed to choose between items of information representing the % of pots made on the two islands which had a given characteristic (e.g., the % of pots made on one island which had curved handles and the % of pots on the other island which had curved handles). These two pieces of information are equivalent to P(D_1/H) and P(D_1/not-H) where H and not-H represent the two islands. The information was presented in a 2 x 6 array (two islands by six pot characteristics) and subjects were told that they could choose any six items from the array. Given this constraint, the normative solution is to choose any three pairs of items; that is, to choose P(D/H) and P(D/not-H)

for three pot characteristics. Out of 121 subjects 11 (9%) chose three pairs; 71 (59%) choose no pairs. Thus the majority never selected information which would allow them to determine either the diagnosticity of the data they had (the pot characteristics) or the posterior odds that the pot came from a given island, and fewer than one in ten selected items in an optimal way. Subjects nevertheless revised their estimates of the likelihood that the pot came from a given island, even though most of them had selected diagnostically worthless information. Doherty, et al. labeled this phenomenon "pseudodiagnosticity".

In all of the research cited above, subjects were attempting to determine the relative likelihood of some set of hypotheses, e.g., is it more probable that a pot came from this island or from that island; does a person have this disease or that disease? These tasks are a subset of a large class of real world problems, for instance a scientist testing two rival theories or a mechanic attempting to determine whether or not the ignition system on a motorcycle is faulty. We will call such problems "Inferences" since they involve the attempt on the part of the person to infer something about the state of the world. The present research also addresses inferences, but there is no implication that the results are generalizable to cognitive tasks other than inferences. Current research in our laboratories suggest strongly that the conclusions drawn above are specific to inferences, a very important class of cognitive tasks.

An attractive feature of the pseudodiagnosticity paradigm is that there is a normatively correct data selection. An example of an inference would be a decision about whether an unknown car is a Toyota Tercel or a Ford Tempo. Assume that two types of information are potentially available about each alternative - the percentage of Tercels and Tempos

which get better than 25 miles per gallon and the percentage of Tercels and Tempos which are driven 50,000 miles with no significant mechanical problems. The problem can be represented in the following manner:

Information Categories		Alternatives 1 Tercel	Alternatives 2 Tempo
1	% over 25 mpg	A	B
2	% over 50,000 miles with no problems	C	D

Assume that a subject is told that the unknown car gets over 25 mpg and that it has gone over 50,000 miles with no mechanical problems; that is, the subject is given two pieces of data, D_1 and D_2. The subject is also given one of the cell entries, say $P(D_1/H_1)$; i.e., cell A. The task is to infer whether the car is a Tercel or a Tempo, and any one of the three remaining probabilities can be chosen. In such a static, two-alternative inference problem where the available data are conditionally independendent, where Cell A is given, and only one more datum may be selected, then:

$$P(H_1/D_1) = \frac{P(H_1)\,P(D_1/H_1)}{P(H_1)\,P(D_1/H_1) + P(H_2)\,P(D_1/H_2)}$$

As noted above, only $P(D_1/H_2)$, that is, Cell B, permits computation of the posterior probability $P(H_1/D_1)$

Note that if this were a problem calling for the person to decide whether to purchase one of the cars, i.e., calling for an action, there is, strictly speaking, no normatively correct set of cell choices. In the case of inferences there is a "true" state of nature and certain cell choices maximize the likelihood of finding out what it is. In the case of actions, however, there is no objective, external criterion ("what is") against which to evaluate a decision.

B. RESEARCH CONDUCTED UNDER THIS CONTRACT

It should be noted at the outset that the tasks set for subjects in this part of the final report use what the introduction calls "pre-encoded" error, i.e., the % values correspond to P(D/H) and P(D/not-H). The error is, in effect, in the relation between the data and the hypothesis, rather than being identifiable as in the data per se. This paradigm is included because it affords an opportunity to determine whether subjects will seek data which bear an unambiguous relation to some hypothesis when the unambiguous data are potentially disconfirming, when potentially confirming but ambiguous data are equally easily available. Hence it is relevant to the general issue of subjects' data selection biases in the face of uncertainty.

Three variations of a single experiment were run. There were three separate versions in order to assure a modest degree of cross-task generalizability within the pseudodiagnosticity paradigm. That is, we wished to rule out the possibility that whatever results were obtained would be attributable to the specific numerical values of the terms in the scenarios.

EXPERIMENT 9a

Subjects. Seventy-two students enrolled in introductory psychology classes participated in the experiment.

Materials and Procedure. All instructions and manipulations were accomplished by varying the content of a two-page booklet, patterned after the ones used in Doherty et al. (1979, 1981). The following text appeared on the first page as an introduction to the task:

"An archaeological expedition is trying to recover artifacts from many ancient, island civilizations. Currently the expedition is recovering artifacts from Emerald Island and Azure Island. Both islands are known to have produced much fine pottery, and both are known to have lost much of this pottery at sea.

Imagine that you are part of this expedition. Specifically, you are a museum curator whose job it is to identify pots brought up from undersea dives. You know the following.

The two civilizations, Azure and Emerald Islands, had kept careful records of all pots shipped out on their merchant ships. These records were uncovered by a previous expedition. Emerald Island is known to have lost 1600 pots at sea, while Azure Island is known to have lost fewer, in fact, about 1200.

One day a call comes in from a wireless on the ship. A new artifact has been found, in perfect condition and it is a pot. You wish to find out as soon as possible which culture has produced the pot, but you are unable to go to the expedition site. The crew on the ship is very busy, but they will speak with you for a short time. You can only ask one question about the artifact at a time when you radio them. You check the merchant's records and find that one way in which pots made by the two islands were different was the type of clay, as follows:

	Azure	Emerald
Red clay	52%	88%
Tan clay	48%	12%

In half of the booklets, the base rate favored Emerald while the remaining booklets favored Azure. The next paragraph indicated that a phone call could be made to determine whether the pot was made of red or tan clay. The phone call was simulated by peeling off an opaque sticker from the answer to one of the following questions:

Is the pot made of red clay?

Is the pot made of tan clay?

Both answers to the 'phone calls' favored the island with the highest base rate. Subjects then were asked to state their hypothesis about which island produced the piece of jewelry.

Page 2 of the booklet instructed Ss that ancient records showed that the two islands tended to specialize in pots of different size. One form of the following table was then printed:

	Emerald	Azure
Small	0%	32%
Medium	23%	46%
Large	77%	22%

	Emerald	Azure
Small	20%	32%
Medium	23%	46%
Large	57%	22%

Ss then made another phone call to determine the size of the pot. The phone call was simulated by peeling off an opaque sticker from the answer to one of the following questions:

Does the pot hold an ounce?

Does the pot hold about 2 quarts?

Does the pot hold several gallons?

Subjects were then asked to state their final conclusion about the source of the pot, but the data about their conclusions are not germane to the issue of data selection.

On page 2, the percentages associated with the pot sizes were manipulated in two ways: (1) a "0" or "23" appeared in column 1, and (2) the "0" or "23" appeared in the Small, Medium, or Large column. These manipulations were crossed with the base rate manipulation (favored Emerald or Azure) to construct twelve forms of the booklet. Six forms assigned the higher base rate to Emerald and used "0" or "23" in each of three pot sizes. Another six forms favored Azure and also used "0" or "23" in each of the three pot sizes. The phone call responses on page 2 were not of interest, nor was the final response: these were on the page only to complete the task for the subjects, since our interest was in the data selection.

EXPERIMENT 9b

Subjects. Seventy-two students enrolled in introductory psychology classes participated in the experiment.

Materials and Procedure. Experiment 9b used modified versions of the booklets described in Experiment 9a. The booklet in Experiment 9b described civilizations on Jasper and Amber islands. Three factors were changed from Experiment 9a: (1) base rates of 1000 and 800 were used, (2) the differences in within-row percentages in the table on page were reduced by 28 (i.e., 1st row= 42% and 39%, 2nd row= 58% and 66%), and (3) a "7" was used in the table on page 2 rather than a "23". Other entries in the table were modified to maintain a sum of 100% within each column.

EXPERIMENT 9c

Subjects. Seventy-two students enrolled in introductory psychology

classes participated in the experiment.

Materials and Procedure. Experiment 9c used modified versions of the booklets used in Experiments 9a and b. The booklet in Experiment 9c described civilizations on the Granite and Quartz islands. Three factors were changed from Experiment 2 to Experiment 3: (1) base rates of 550 and 450 were used, (2) the differences in within-row percentages in the table on page 1 were reduced by 2 (i.e., 1st row= 54% and 48%, 2nd row= 46% and 52%), and (3) a "13" was used in the table on page 2 rather than a "23". Other entries in the table were modified to maintain a sum of 100% within each column.

RESULTS

The results of the three experiments, 9a, b and c, will be treated together, since they are designed to be replications of one another. The great majority of the subjects responded appropriately to the task, 92% of them concluding that the island favored by the base rate and by the data on page was the favored island. The data choices made by each subject on page 1 were tallied. Table 3-1 shows these choices broken down by the island favored by the base rate, and totaled over islands by whether the choice of datum about which to ask would, assuming a "yes" response, tend to favor the island already favored by the base rate. Considering the favored island as a hypothesis under test, the frequency tally described is tantamount to a test for confirmatory bias. (See Table 3-1). The frequencies shown in bold face are those consistent with a bias to confirm the hypothesis that the base rate favored.

Table 3-1. The frequency of data choices[a] associated with each set of P(D/H) values on page 1.

Island favored by the base rate	feature	Data set		Choice frequency
		Jasper	Amber	
Amber	Loose lid	42 %	34 %	13
(1000:800)	Attached lid[b]	58%	66 %	**23**
		Amber	Jasper	
Jasper	Loose lid	42 %	34 %	**23**
(1000:800)	Attached lid[b]	58%	66 %	13
		Quartz	Granite	
Quartz	Made of gold[b]	54 %	48 %	**28**
(550:450)	Made of silver	46 %	52 %	8
		Granite	Quartz	
Granite	Made of gold[b]	54 %	48 %	**33**
(550:450)	Made of silver	46 %	52 %	3
		Azure	Emerald	
Emerald	Red clay[b]	52 %	88 %	**16**
(1600:1200)	Tan clay	48 %	12 %	20
		Emerald	Azure	
Azure	Red clay[b]	52 %	88 %	18
(1600:1200)	Tan clay	48 %	12 %	**18**
Totals		Favoring base rate		**141**
		Counter to base rate		75

[a] The frequencies in bold face are of those choices favoring the base rate.
[b] the feature which is present in the pot on page 1.

Note that there is a strong tendency in the page 1 choices for subjects to prefer data likely to favor the hypothesis (such frequencies are shown in bold face) they already hold. This tendency shows up in the choices of those subjects in the two versions of the experiment in which the data were relatively non-diagnostic (note that the P(D/H) ratios were relatively similar for the data in both the Quartz/Granite and the Amber/Jasper comparisons. In the the Emerald/Azure comparison, the 4:1 ratio for tan clay apparently influenced a number of subjects to select tan clay about which to ask. Even with the data of the Emerald/Azure comparison included, the overall χ^2 for the entire data set (141 vs. 75) was significant, ($\chi^2(1) = 20.17$) supporting the hypothesis that confirmatory bias influences data selection.

The data choices on p. 2 were tallied in a fashion similar to that for page 1. Table 3-2 summarizes these choices, again showing the choices favoring the base rate in bold face. Since the basic purpose of this investigation was to determine if subjects would seek out data that had the potential for being perfectly diagnostic, i.e., data pairs where one of the values of the % was 0, the data are further broken down by whether the set of choices available to the subject did or did not include a pair with a 0% as one of the three pairs of %s. We will refer to the datum associated with the 0% as the "0% datum". The corresponding datum in the matched set (i.e., the datum in Amber or Jasper associated with 11%, the datum in Quartz or Granite with 13%, and that in Emerald or Azure with 32%) will be referred to as the "designated datum".

Table 3-2. The frequency of data choices[a] associated with each set of P(D/H) values on page 2, collapsed over the position of P(D/H). The choices data are presented separately for those forms which had a zero P(D/H) and for those that did not.

Island	Data type	P(D/H) Values		
		0 & 11 or 7 & 11	53 & 27 or 46 & 27	47 & 62
Amber	0%	8	**8**	2
	11%	3	**8**	7
Jasper	0%	**8**	4	**6**
	11%	**1**	9	**8**
		0 & 22 or 13 & 22	56 & 21 or 43 & 21	44 & 57
Quartz	0%	**8**	1	**9**
	13%	**1**	5	**12**
Granite	0%	7	**3**	8
	13%	1	**7**	10
		0 & 32 or 20 & 32	77 & 22 or 57 & 22	23 & 46
Emerald	0%	10	**6**	2
	32%	3	**12**	3
Azure	0%	7	7	**4**
	32%	**1**	10	**7**

[a] The frequencies in bold face are of those choices favoring the base rate.

While these are virtually raw data, close inspection shows that the presence of a 0% attracts subjects to select the datum associated with that 0% more frequently than the designated datum. Table 3-3 sums the frequencies over the three experiments. This gives a clearer picture of the influence of the 0% datum, which is potentially perfectly diagnostic.

Table 3-3. The data of Table 3-2 collapsed over island names and values of P(D/H), sorted into choices more likely to confirm or to disconfirm.

Data type	0% datum Designated datum	Confirming	Disconfirming
with 0%	**23** + 25	**36**	24
without 0%	**3** + 7	**54**	44
Σ	**26** 32	**90**	68

Note that those choices associates with the 0% datum are as likely to be for the 0% when it is disconfirming as when it is confirming. Therefore, we collapsed the frequencies in the first column of data in Table 3-3 to obtain Table 3-4.

Table 3-4. The data of Table 3-3 with the data selection frequencies associated with the 0% datum collapsed over confirmatory and disconfirmatory selections.

Data type	0% Datum Desig. Datum	Confirming	Disconfirming
with 0%	48	36	24
without 0%	10	54	44
Σ	58	90	68

A χ^2 test for association on the data of Table 3-4 shows that the subjects' data choices are influenced jointly by the presence of the 0% datum and by whether the nonzero data are confirmatory or disconfirmatory: $\chi^2(2) = 34.38$ ($p < .01$). A χ^2 test of the null hypothesis that the designated datum is as likely to be chosen as the 0% datum (48 vs. 10) suggests that the 0% datum does indeed get chosen more frequently ($\chi^2(1) = 24.90$, $p < .01$) than its counterpart. Note that the tendency of subjects to select the 0% datum is not influenced by the magnitude of the % associated with the designated datum: $\chi^2(2) = 0.12$. n.s., (see Table 3-5).

Table 3-5. Frequency of choice of the 0% datum as a function of the % associated with the designated datum.

% associated with the designated datum	11%	13%	32%
Frequency of choice of 0% datum	16	15	17

Finally, note that the normally robust finding of a bias to confirm seems to have been disrupted by whatever cognitive processes have been engaged by the the time that the subjects are selecting the data on page 2. There no tendency toward a bias to confirm in the data selections of the pairs that include a 0% datum. Nor is there a significant tendency toward a confirmatory bias on the part of those subjects who chose from the other pairs of %s, the pairs that included neither a % datum nor a designated datum (90 vs. 68 from Table 3-4): $\chi^2(1) = 3.06$, n.s.

While the normally robust tendency toward a confirmatory bias may have been disrupted by the presence of the 0% datum, it may also have been

attenuated by an unexpected phenomenon, which to our knowledge has not been observed before. There was a strong tendency for subjects to select large percentages, irrespective of their diagnostic impact. There are 30 pairs of %s listed in Table 3-2, exclusive of those with a 0% datum, and the correlation between the sum of the %s for each such pair and the frequency of data selection for that pair is $r(28) = .58$, $p < .01$. This is a nonrational phenomenon, and cannot be explained by positing a relationship between the magnitudes of the numbers and a confirming vs. falsifying relation to the hypothesis under consideration: the design of the study makes that relationship precisely zero. Nor is there in this study a relationship between diagnosticity and magnitude. This unforeseen phenomenon apparently obscured the expected effect of a bias to confirm, expected to appear in data choices other than those of the 0% datum.

DISCUSSION

This set of studies shows clearly that some subjects are sensitive to diagnosticity, when that diagnosticity is made highly salient. The data of page 1 indicated that the great majority of subjects (74%) sought confirmatory evidence when the data were not highly diagnostic. That dropped to 47% when the disconfirming datum had a likelihood ratio of 4:1, as opposed to the 1.69:1 likelihood ratio for the potentially confirming datum. On page 2, the subjects were almost 5 times as likely to select a 0% datum as they were to select a designated datum. This is a powerful effect, except that we must bear in mind that it was still a minority (48 of 108, or 44%, see the top row of Table 3-4) who chose the 0% datum.

As noted in the introduction to Part 3, this study does not manipulate data error in the sense described in the general introduction. But this study does show clearly that subjects are sensitive to the possibilities

associated with data that may be, in a sense, perfect. Many subjects select such data even when the data are likely to provide evidence against their hypotheses. Diagnosticity effects appear, however, only when the diagnosticity is extreme. Moderate differences in diagnosticity do not influence data selection at all, as the two right hand data columns in Table 3-2 indicate.

While it was not part of the formal proposal, an additional study was conducted to show the power of confirmatory bias in cognitive tasks, and to try to gain some insight into an explanation of that confirmatory bias.

EXPERIMENT 10a

Subjects. One hundred thirty five students enrolled in the introductory psychology course served as subjects.

Materials and Procedure. All instructions and questions were presented in a booklet. The following text was presented:

> Suppose you were shown two large sacks, one called A, the other B. Suppose further that you were told that both sacks were full of red and blue marbles, in different proportions. Sack A is filled with thousands of marbles, 70% of them are red, 30% are blue. In Sack B there are also thousands of marbles, but the proportions are reversed, with 70% of the marbles being blue and 30% being red.
>
> I reach into the sack on the right and draw out a handful of 12 marbles. There are 7 red marbles and 5 blue ones in the sample. This is fairly strong evidence that the sack on the right is Sack A, But there is some chance that it is the other sack. If you had to decide whether I had sampled from Sack A or from Sack B, and I let you draw another sample, that is, a second handful of marbles, would you draw that sample from the sack on the right or from the sack on the left?
>
> The left sack _____ The right sack _____

RESULTS

The sack named in the version as the one from which the sample had been drawn was randomized. The position of the choices was always as shown. A Bayesian analysis shows that the information provided by either choice is equally informative. Subjects strongly preferred to sample the same sack that had been named in the problem statement (i.e., the right sack in the above version). The effect was, to put it mildly, strong, with 121 preferring the same sack and 14 preferring the alternate one: $\chi^2(1) = 64.00$, $p < .01$).

EXPERIMENT 10b

Subjects. One hundred sixteen students enrolled in the introductory psychology course served as subjects.

Materials and Procedure. All instructions and questions were presented in a booklet. The following text was presented:

> Suppose you have two decks of cards in front of you. They have been mixed up so that one deck has 75% black cards and 25% red cards, instead of the usual 50/50 split. The other has 25% black and 75% red. Your task is to tell me which deck is which, by observing two cards. I turn over a card from the deck on your right and it is red. That is fairly strong evidence that the right hand deck is the predominantly red deck, but there is some chance that it is not. Now I am going to turn over another card. From which deck would you want me to turn over the second card?
>
> The left hand deck ____ The right hand deck ____

RESULTS

As in the sack task, the deck named in the statement was randomized, but the choices were as shown. Again the preference was extreme, with 105 subjects preferring to draw a card from the same deck, and 11 preferring the other deck: $\chi^2(1) = 76.17$, $p < .01$. Across the two versions

of this task, fully 90% of the subjects preferred to sample from the same data source that they already had data about. Combining the two groups of subjects, we see that 226 preferred the same source, 25 the other source: $\chi^2(1) = 160.96, p < .01$.

DISCUSSION

Both version of this data selection task shows show a powerful bias to select information about a single hypothesis. Note that there does not seem to be any way to explain this as a "bias to confirm," given the structure of the task. We interpret this as a possible, nonmotivational explanation of the widely cited "confirmatory bias" (Tweney, Doherty & Mynatt, 1981), since tasks in which subjects show confirmatory bias also have the quality of having subjects select data about a single hypothesis (Doherty & Mynatt, 1987). We will return to this in the general discussion, but we believe that this same inability (or unwillingness) to consider alternate hypotheses may underlie the destructive effects of SF error.

PART 4

RESEARCH ON DATA ERROR IN THE ARTIFICIAL UNIVERSE TASK

CONTENTS

A. Prior Research.

In studies of error effects many decisions must be made concerning how error is introduced into the task. One decision concerns whether the error is located in the input data which form the basis for inferences or predictions, or in the feedback concerning their correctness. Experiments 11 and 12 address this issue and, in addition, seek to generalize the findings of the earlier studies to more complex task domains.

As noted in the introduction, Kern (1982) first distinguished between system failure (SF) and measurement (ME) error. Her study began with the role of confirmation bias in scientific inference, and asked whether the error in data could serve subjects as a source of auxiliary assumptions used to protect a favored hypothesis from disconfirmation.

The task used by Kern was based loosely on "artificial universe" studies of scientific inference (Mynatt, Doherty, & Tweney, 1977;1978). Subjects were asked to imagine that they were scientists investigating an unexplored planet from an orbiting spaceship. By dropping probes to the planetary surface, they could determine which regions of the planet were capable of supporting life. Subjects were asked to determine a line on the planet surface which separated regions capable of life-support from regions incapable of life-support. On each trial, subjects could drop a probe to the surface at a region of their choice and determine whether imaginary plants (called "tribbles") survived or failed to survive. Based on data from the probes, subjects were asked to move a hypothesized boundary line to a position that distinguished the two regions. Kern's display screen is shown in Figure 4-1.

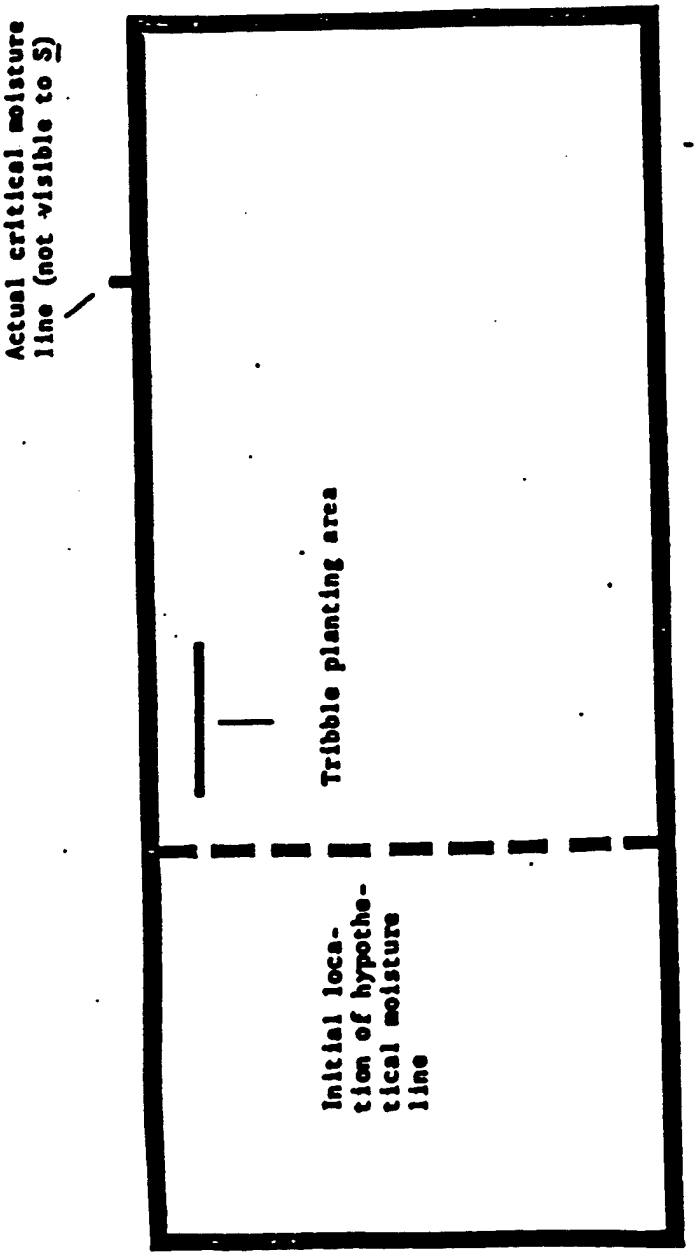


Figure 1: Illustration of initial display screen. The investigator indicated to S that the large rectangular area represented the site of the investigation.

Both SF error and ME error were manipulated, but in very different ways. For ME, subjects were told that the locations of dropped probes on the planet surface were known only within certain limits represented by the horizontal bar in Figure 4-1. The length of the bar corresponded to the positional uncertainty of the probe. For SF, subjects were told that the telemetry device in the probe (which indicated whether the tribbles lived or died) could fail on 25% of the trials. Thus, whether the device reported "lived" or "died" was determined randomly on 25% of the trials, rather than by the actual location of the tribble relative to the critical line.

Kern found that subjects faced with ME were no different in accuracy of placement of the final hypothesis line than control subjects whose data contained no ME. Similarly, in experiments 1-6, described earlier in the present report, subjects seemed able to "average" over such error. However, subjects in Kern's study who had to cope with SF error did much worse than control subjects given accurate feedback. In addition, she found that subjects were more likely to challenge the accuracy of probes (through a "probe-check" routine) when the feedback disconfirmed their hypothesis. Thus, the presence of SF error apparently led subjects to invoke error to "explain away" data not confirming their current hypothesis. Kern's results thus reflect outcomes similar to our results in Part 1, experiments 1 through 6.

Kern's research was innovative and insightful, but interpretation of her data is made difficult by the presence of inadvertent confounding in her design. By placing ME in the location of the dropped probes, and SF on the feedback from the probes, Kern confounded type of error with locus of error. Could this make a difference?

B. RESEARCH CONDUCTED UNDER THIS CONTRACT

Experiments 11 and 12 separately manipulated SF and ME at each of the two loci of error. Experiment 11 examined the effects of ME and SF located in the input, and 12 the effects of ME and SF in the feedback. Since survival was causally dependent upon location, input was defined as data which reflected the location of dropped probes. Feedback was defined as data which reflected the survival of the tribbles. In experiment 11, error of either type could affect the ability of subjects to drop a probe at a predetermined location, but not the validity of results from each probe; whatever the location of the probe, the feedback (whether tribbles lived or died) was veridical. In experiment 12, error was added to the feedback from the probe, but did not affect a subject's ability to place a probe at a predetermined location. Thus, probes always landed at the subject's choice of location, but error of either type was potentially present in the feedback whether the tribbles lived or died.

The task environment and procedure were similar to those used by Kern (1982), in that subjects in both experiments 11 and 12 were asked to "role play" the part of a scientist investigating an unexplored planet. Specifically, subjects were asked to launch tribbles to the planet's surface from a spaceship visible on the screen. Subjects were told that survival of tribbles depended only on the moisture content present in the planet's soil; tribbles would grow above a certain moisture content, below this moisture content they would die. Subjects were asked to determine the critical moisture level by launching probes to various points on the planet's surface to see whether the tribbles survived at each location.

To facilitate the task, subjects were told that the percentage of soil moisture on the planet surface increased uniformly from west (the left side of the screen) to east (the right side of the screen). A vertical line across the planet surface, indicating a tentative hypothesis for the minimum moisture level necessary for the tribble colony to survive, was shown on the screen at the beginning of the task. Before each launch, subjects were asked to predict whether the tribble they were about to launch would live or die. After each launch, subjects were given a chance to move the moisture level hypothesis line to a new position.

Experiment 11

METHOD

Subjects. The subjects in this experiment were 63 graduate and undergraduate students, each of whom was paid $5.00 for participating. Subjects were recruited from classes in introductory and abnormal psychology, honors psychology, biology, chemistry, and geology. Subjects were randomly assigned to experimental conditions. Three levels of ME error: None: 2 units (I. e., the tribble was represented as landing in an area that was 2 pixels wide), Low: 20 units, High: 80 units. ME was crossed with three levels of SF error (None: 0%, Low: 25%, High: 50%) yielding nine experimental conditions with seven subjects in each cell.

Procedure. Subjects signed up to participate in groups of 2 to 4 persons each. After meeting the experimenter in a waiting area, subjects were taken to a small computer lab. Four Macintosh computers that had been set up for the appropriate experimental conditions prior to subjects' arrival displayed the beginning screen as subjects entered. Before the task began,

subjects were told that they were all working on variations of the same task and that the computers would make a variety of sounds. Brief instructions were given on the use of the mouse to select commands for the computer. Subjects were asked to read through the instruction booklet next to their computer and told that they could ask questions at any time. Subjects began the task when they finished reading the instructions.

Each subject launched a series of eight tribbles, one at a time, and received feedback after each launch about whether the tribble lived or died. The actual probability of survival of a tribble launched to a specific location was either 0.0 or 1.0, determined by whether the launch was the left or right, respectively, of the critical line. After the completion of eight launches, subjects in all conditions were informed that enough resources were still available for four more tribbles to be launched. Thus, all subjects launched a total of 12 tribbles.

ME error was manipulated by varying the width of the rectangle shown on the screen representing the approximate location on the planet's surface to which the tribble had been launched. In the High and No SF error conditions, the instructions emphasized that "the planting area depicted on the computer screen is only the area where the tribble was **intended** to be launched to, not necessarily the area where it was actually launched to". Pages 158 - 169 provide an example of the instructions, which varied somewhat depending on the experimental condition. These are specifically the instructions from the Low ME- No SF condition.

TRIBBLES EXPERIMENT

****
* INSTRUCTIONS *
****

You are a scientist investigating an unexplored planet, Ethereus. Right now, you are orbiting Ethereus in a spaceship. From your spaceship, you can conduct a variety of controlled experiments. Previous research has shown that certain life forms exist on the planet, but the conditions which support these life forms are very poorly understood. Your research project will involve the growth of a plant, the tribble, found in certain regions of Ethereus. The tribble was selected as the focus of this initial investigation because earlier work suggests that its survival depends only on the amount of moisture present in the soil. It is suspected that above a certain moisture content, tribbles grow. Below this moisture content, the tribbles die.

****

Your task is to determine what this critical level is by systematically planting tribbles at various points on the planet's surface and seeing whether or not they survive at these locations. Each of the points you select for planting will have a certain moisture level, which determine whether the tribble lives or dies there.

****

The site of the investigation will be a 250,000 square-mile area encompassing a large portion of the planet's southern hemisphere. Referring to Figure 1 (a picture of what the computer screen will show during the experiment), this area of investigation is depicted by the large box encompassing the bottom 2/3 of the figure (see A).

Refer to Figure 1

Fortunately, research has established that the distribution of moisture in the planet's soil is remarkably regular. The percentage of soil moisture on the planet's surface INCREASES uniformly from west (left side) to east (right side). Preliminary data indicate that tribbles can survive only when the percentage of moisture in the soil equals or exceeds a certain level. This level is indicated by the solid line drawn vertically across the planet's surface. (See B). Your job is to conduct a more thorough investigation of the tentative hypothesis that tribbles can survive only at moisture levels equal to or exceeding the level displayed by this line, by planting them east or west of the line and observing whether they live or die.

Each time you are ready to plant a tribble, you will see your spaceship appear in the orbit-area above the planet (See C). At this time, you are ready to prepare to launch a tribble to the planet's surface. Position your spaceship launcher (the marker protruding from the bottom of the spaceship)

exactly above the place you wish to plant the tribble. Positioning is conducted by the spaceship control panel in the upper right corner of your computer screen (See D). To activate the spaceship, you move the cursor/arrow (by moving the mouse) to the desired button, position the arrow within the button (as is shown on the figure), and then click the mouse once to activate the button. The arrow buttons designate the direction of spaceship movement. If you want to move the spaceship to the East, use the "---->" button. If you want to move the spaceship to the West, use the "<----" button. The "STOP" button stops the spaceship from moving in any direction. YOU MUST STOP THE SPACESHIP FROM MOVING IN ANY DIRECTION BEFORE YOU CAN EXECUTE ANY OTHER CONTROL PANEL COMMAND (SUCH AS, MOVING IN THE OPPOSITE DIRECTION, OR LAUNCHING A TRIBBLE).

Once the spaceship launcher is pointed directly above the place you wish to launch a tribble, begin the launching process by clicking the "LAUNCH" button on your control panel. Your spaceship's on-board computer controls the execution of the launching procedure.

Once the tribble has been launched, a rectangle will appear on the screen under the spaceships' launcher. This rectangle represents the planting area at which you positioned your launcher to drop the tribble. In other words, where you intended the tribble to be planted.

*****See Figure 2*****

Why is the planting area represented by a rectangle instead of a single point? Because you are orbiting the planet from a distance of 500 miles, you are not able to drop tribbles at a precise location. Rather, tribbles land within 15 miles of the location at which you aim. Although the tribble may land somewhat east, west, north, or south of the location you specify, only east-west error is of interest with respect to your moisture hypothesis, so this is the error you see represented on the screen by the rectangular box. Moving this rectangle east or west on each trial specifies a planting location. When the tribble is released, you will know that it has landed somewhere within the east-west range represented by the rectangle, but you can be no more specific in your observations than this.

In order to determine where the actual moisture line lies, it is important to determine if the tribble you planted lived or died at that planting location. So, every time you launch a tribble, one telemetry device is sent down with it. The telemetry device is located wherever the tribble is planted. The device beams back to your ship, almost immediately, whether the tribble lived or died. Previous work with this device indicates that it is 100% accurate. A black tribble on your screen represents a dead tribble. A

white tribble on your screen represents a live tribble.

After you have made the command to launch a tribble, you will be asked to try and predict, based on the data you have available to you, whether the tribble you're about to plant will live or die. Initially, you may feel you don't have enough information to warrant a reasonable prediction, but I'd like you to do your best.

*****See Figure 3 *****

You indicate your prediction by moving the mouse and positioning the arrow on the response button desired. To make your prediction, you then click the button once.

Because the tribble launching procedure is a multi-phase operation involving complex machinery controlled by an on-board computer, it sometimes malfunctions. All five phases of the launch cycle must be very precisely executed, and if any phase is even slightly off specifications, the launcher will over- or under-fire. Over- or under- firing means that the spaceship launcher positioned the tribble to be fired at some area beyond (to the east of) or behind (to the west of) the area you intended the tribble to be launched to. Therefore, if a malfunction occurs in the computer's launching system, you do not know where the tribble actually landed. **Remember, the planting area depicted on the computer screen is only the area where the tribble was **intended** to be launched to, not necessarily the area where it was actually launched to. Therefore, if the computer launching operation malfunctioned, the tribble was launched to some unknown area. However, if the computer launching operation was successfully executed, then the tribble was launched to the planting area intended, which is depicted on the computer screen.

Previous work involving these launching system malfunctions indicated that the rate of malfunction, on the average, occurs 50% of the time. Because this is only an average, your own rate of system malfunction may be slightly greater or less than that.

How can you tell if there was a system malfunction? Well, there is a computer system probe check that you can conduct. Because so much energy and computer memory is involved in executing a probe check, you can only conduct the probe checks twice. The probe check will tell you if the computer launch operation was successful (and you know that the tribble was planted within the area represented on your screen) or was a malfunction (in which case you don't know where the tribble was actually planted).

After you plant a tribble and find out whether it lived or died, you will

be asked whether you want to conduct a launching system probe check. This question will appear in your control panel.

*****See Figure 4*****

Indicate whether you want to conduct a probe check by moving the cursor to the desired response button (YES or NO) and clicking it ONCE.

You have resources available to plant a total of 8 tribbles.

After you have planted a tribble, determined whether it has lived or died, and are preparing to launch your next tribble, a question will appear on the control panel of your computer screen which says "DO YOU WANT TO MOVE THE HYPOTHETICAL LINE?"

*****See Figure 5*****

What I'm interested in finding out is whether or not the data you've generated have led you to reject the location of the original critical moisture line in favor of a new location. I want you to relocate this line only when your data indicate that the line's present position is wrong and does not represent the actual critical moisture level. Again, you respond by positioning the arrow over the desired response button (yes or no) and click it once.

If you decide not to move the line, the computer will give you your next tribble to plant.

If you have decided to move the line, a line-moving control panel will appear, which works the same way you used the spaceship-moving control panel.

*****See Figure 6*****

To move the line left or right use the arrow buttons, and use the STOP button to stop movement in any direction.

When you've reached a point that represents your new working hypothesis about where the critical moisture line should be, click the "FINISH" button once. The line will be repositioned, and the computer will now give you your next tribble to plant.

This entire procedure will be followed for each tribble you plant until you've planted a total of 8 tribbles.

If you have any questions, please ask for clarification from the experimenter before the experiment begins.

To reduce any possible confusion, there is a box in the upper left corner of your screen which keeps account of task information for you. Refer back to one of the figures and note the box. This information is updated each time you are preparing to launch a new tribble.

FIGURE 1

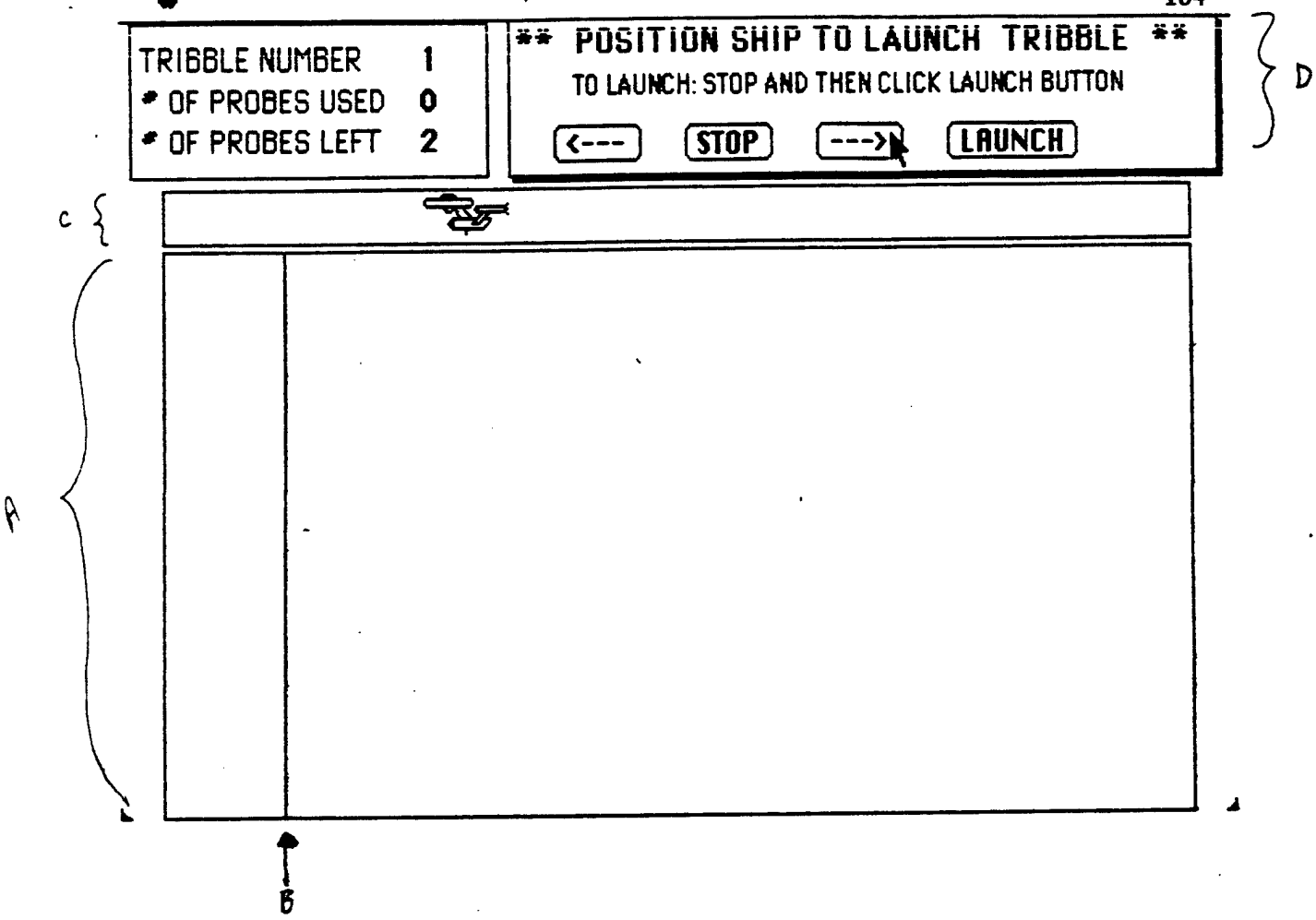

164
TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
** POSITION SHIP TO LAUNCH TRIBBLE **
TO LAUNCH: STOP AND THEN CLICK LAUNCH BUTTON
<---
STOP
--->
LAUNCH
D
C
A
B

FIGURE 2

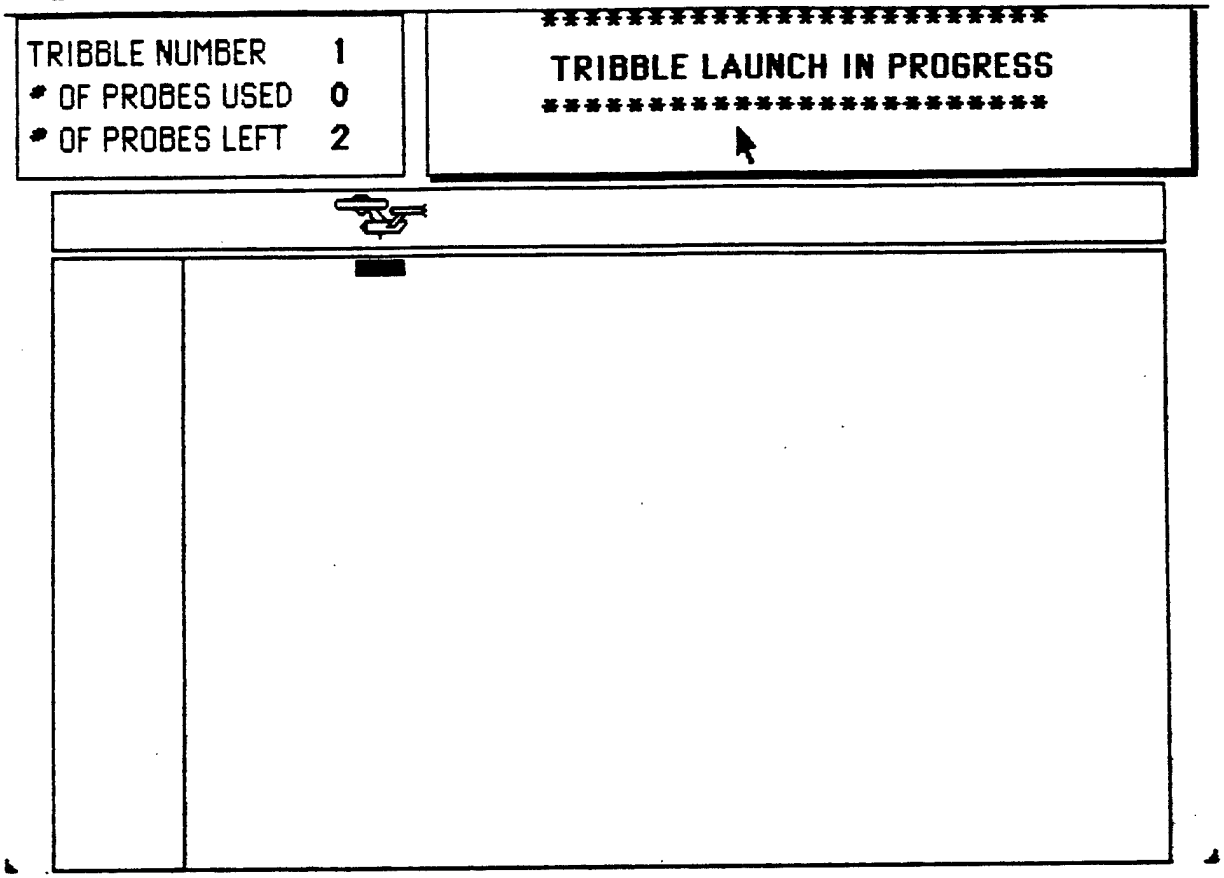

TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
****
TRIBBLE LAUNCH IN PROGRESS
****

FIGURE 3

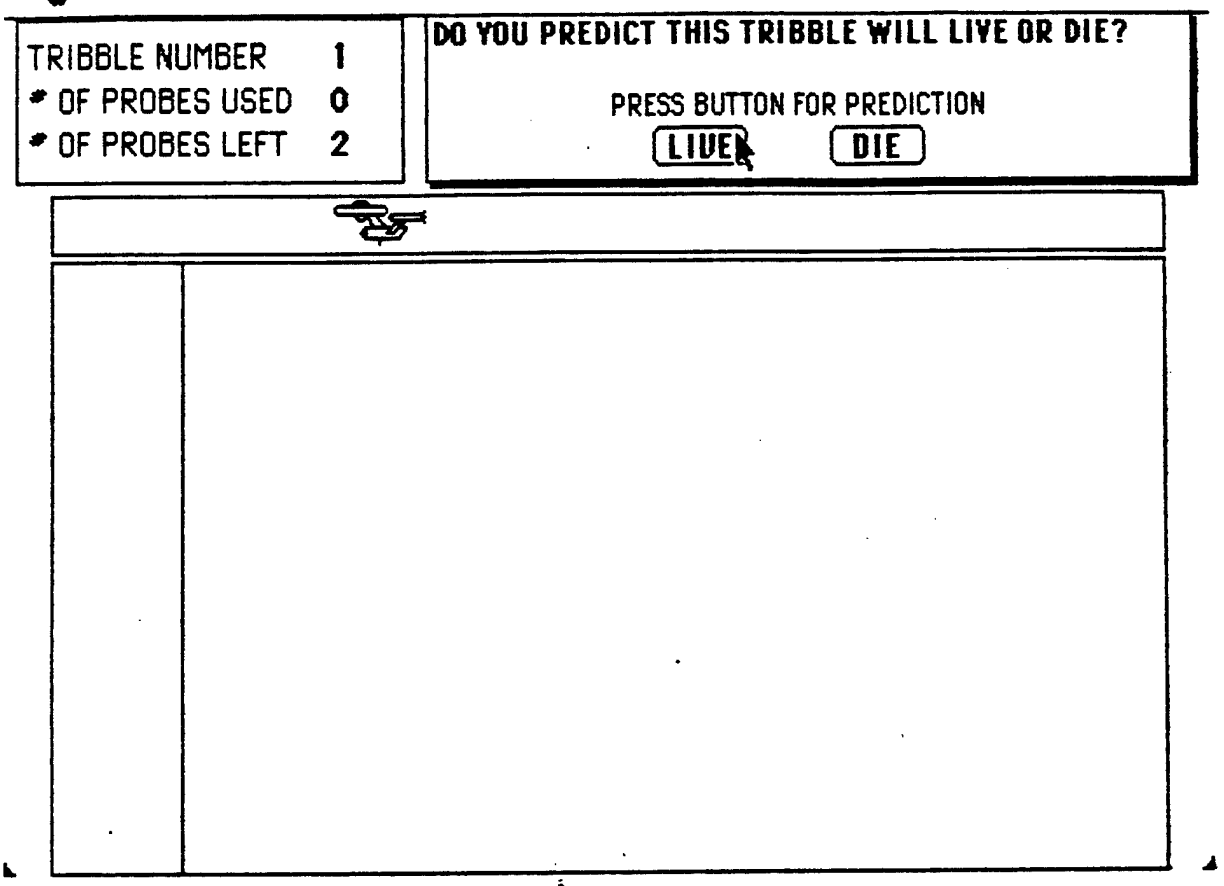

TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
DO YOU PREDICT THIS TRIBBLE WILL LIVE OR DIE?
PRESS BUTTON FOR PREDICTION
LIVE
DIE

FIGURE 4

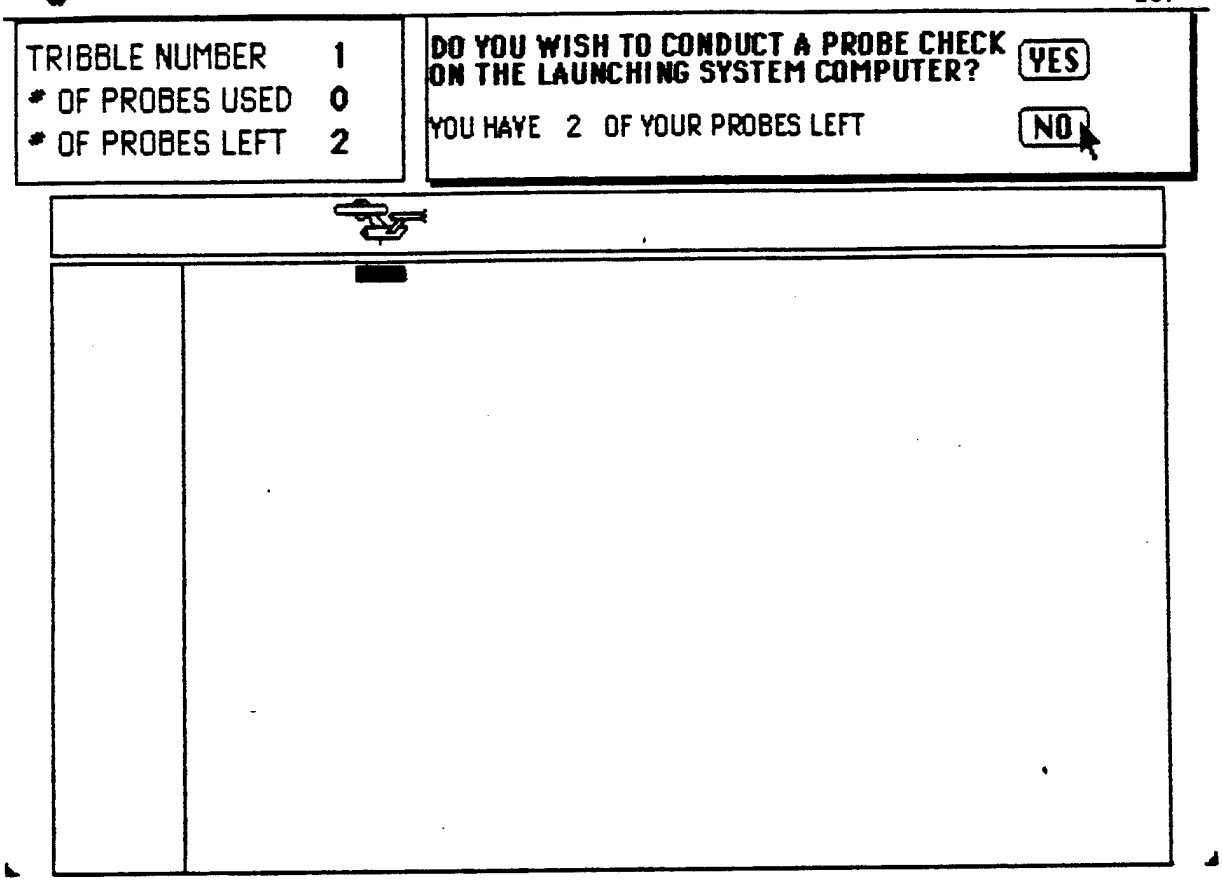

TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
DO YOU WISH TO CONDUCT A PROBE CHECK ON THE LAUNCHING SYSTEM COMPUTER?
YES
YOU HAVE 2 OF YOUR PROBES LEFT
NO

FIGURE 5

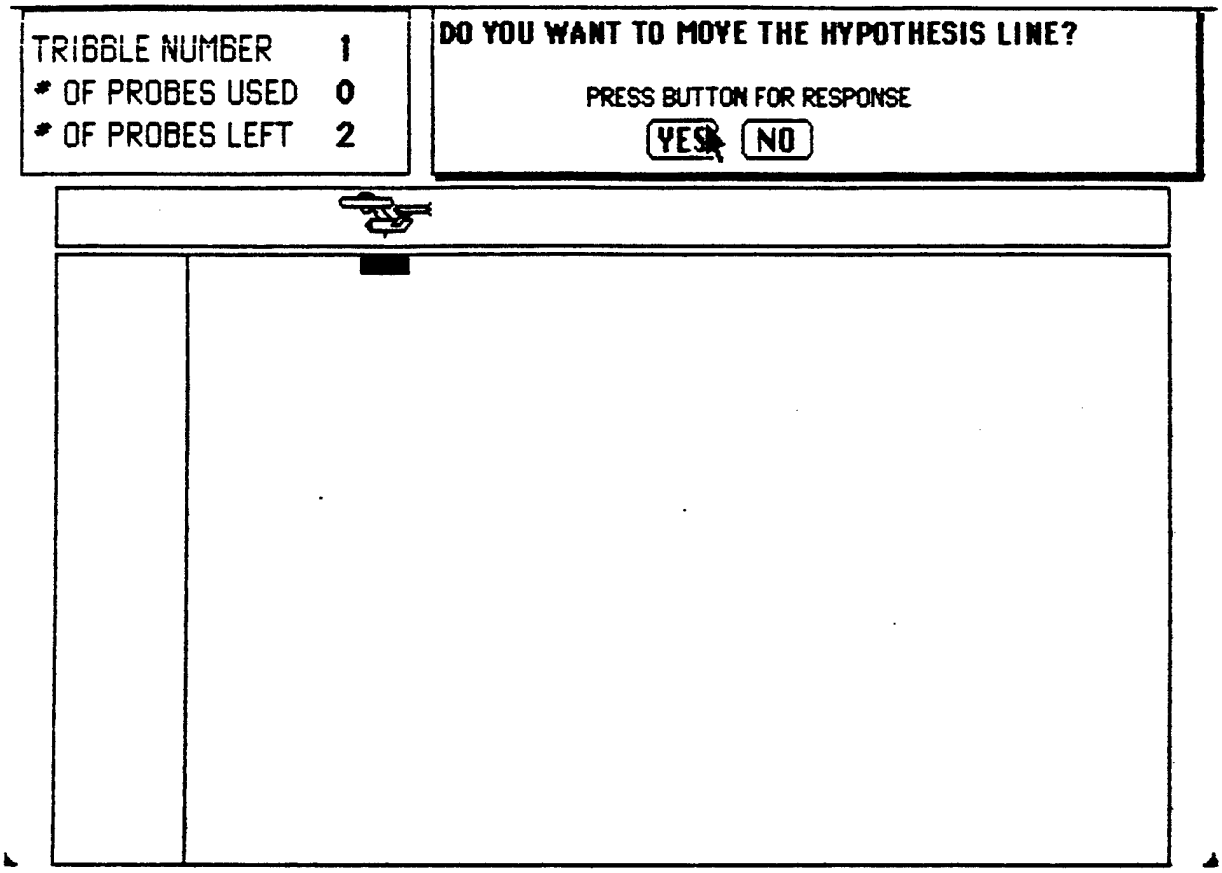
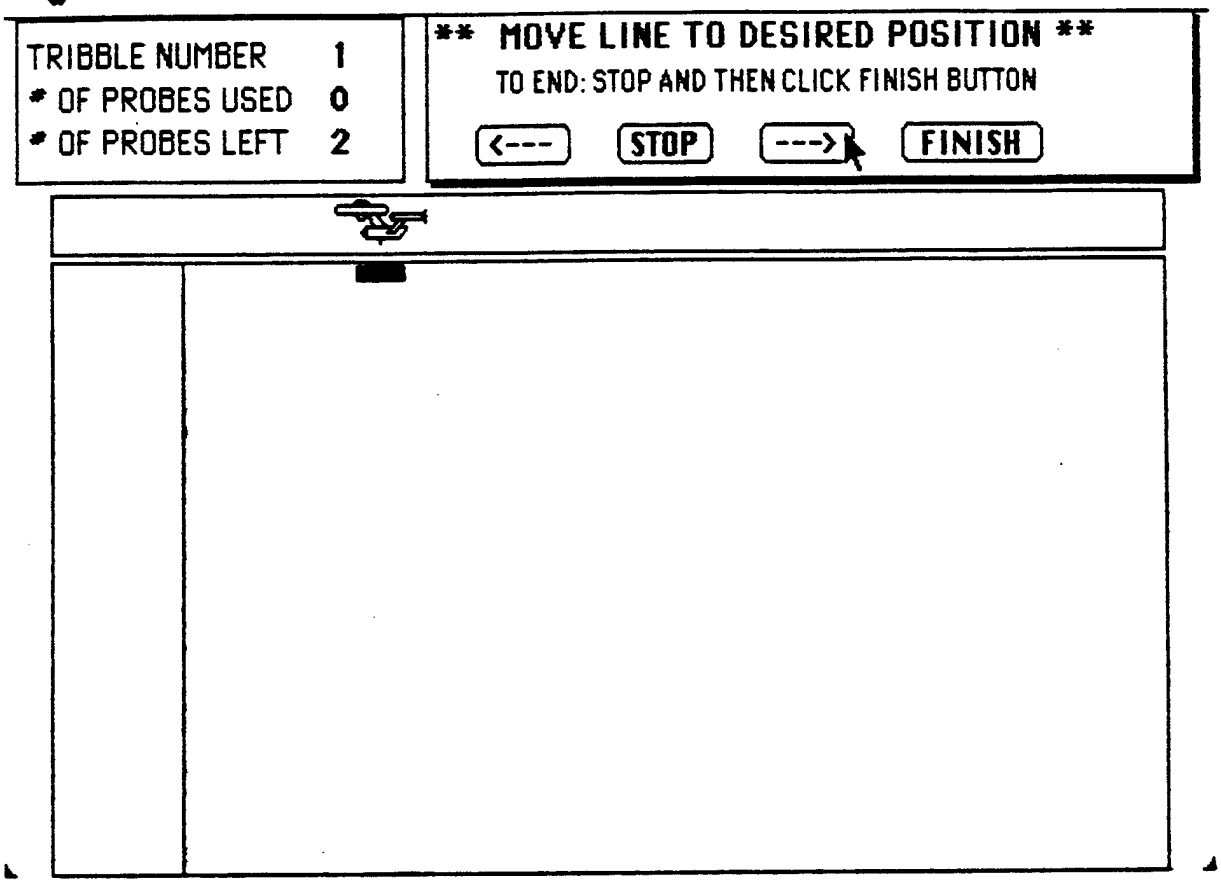

TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
DO YOU WANT TO MOVE THE HYPOTHESIS LINE?
PRESS BUTTON FOR RESPONSE
YES
NO

FIGURE 6


TRIBBLE NUMBER 1
OF PROBES USED 0
OF PROBES LEFT 2
** MOVE LINE TO DESIRED POSITION **
TO END: STOP AND THEN CLICK FINISH BUTTON
<---
STOP
--->
FINISH

Experiment 12

Subjects in this experiment completed a task similar to experiment 11. However, in order to create both ME and SF error that would make sense in the feedback, subjects were instructed that each launch would plant a colony of 600 tribbles. Subjects received feedback about the number of tribbles that lived, presented as an approximate range of values (e.g., they might receive feedback that 320 - 380 tribbles lived). Instructions emphasized that the true number of surviving tribbles could be anywhere between the limits given, and that they should not assume it was in the middle of the range. Note that such limits differ from confidence intervals which locate a sample value in the middle of a reported range. Subjects were informed that at least 450 tribbles must survive in order for the colony to survive.

METHOD

Subjects. The subjects for this experiment were 48 introductory psychology students, half of whom were enrolled in an Honors section. All subjects received course credit for participation. Subjects were arbitrarily assigned to experimental conditions. Each condition represented a specific level of ME error (high, 80 units, or low, 5 units) and a specific level of SF error (high, failure 30% of the time, or none, failure 0% of the time), yielding four experimental conditions: High ME-High SF; Low ME-High SF; High ME-No SF; Low ME-No SF. Because of the sounds generated by the computer involved with probe checks, subjects assigned to the High SF conditions were run separately from subjects participating in the No SF conditions.

Procedure. The procedure used was similar to that described for

experiment 11. Subjects initially launched a series of eight colonies of 600 tribbles each, one colony at a time, and received feedback about the number of tribbles in the colony that survived after each launch. The actual number of tribbles that survived on any one launch was computed using a pre-programmed monotonically increasing function. As in experiment 11, after completion of eight launches, subjects in all conditions were informed that enough resources were still available for four more tribble colonies to be launched. Thus, all subjects launched a total of 12 colonies. The instructions from the High ME- High SF condition are given on pages 172-179.

In the conditions with high ME error (High ME-High SF and High ME-No SF), the feedback that subjects received about the number of tribbles in the colony that lived was computed by adding a randomly generated number between -80 and +80 to the actual number of tribbles that survived. Subjects in low ME error conditions received information computed by addition or subtraction of a random number between -5 and +5. Thus, for example, if 250 tribbles actually lived, this was presented to subjects as an interval either 5 units wide or 80 units wide, with the numerical value given as, say, 249-254 or 220-300, depending on condition. The endpoints were computed randomly, with the constraint that the true value was always contained within the resulting interval (Figure 4-3 shows a typical screen after the first trial). The bars shown on the screen in Figure 4-3 were 20 pixels wide in both ME conditions.

In order to manipulate SF error, subjects in the High ME-High SF and Low ME-High SF conditions were informed that the system used to indicate tribble survival was known to malfunction approximately 30% of the time.

You are a scientist investigating an unexplored planet, Ethereus. Right now, you are orbiting Ethereus in a spaceship. From your spaceship, you can conduct a variety of controlled experiments. Previous research has shown that certain life forms exist on the planet, but the conditions which support these life forms are very poorly understood. Your research project will involve the growth of a plant, the tribble, found in certain regions of Ethereus. The tribble was selected as the focus of this initial investigation because earlier work suggests that its survival depends only on the amount of moisture present in the soil. It is suspected that above a certain moisture content, tribbles grow. Below this moisture content, the tribbles die.

****

Your task is to determine this critical level by systematically planting colonies of tribbles at various points on the planet's surface and seeing whether or not they survive at these locations. Each of the points you select for planting will have a certain moisture level, which determines whether the tribble lives or dies there.

The site of the investigation will be a 250,000 square-mile area encompassing a large portion of the planet's southern hemisphere. In Figure 1 (a picture of what the computer screen will show during the experiment), the area of investigation is depicted by the large box around the bottom 2/3 of the figure (labelled A in Figure 1).

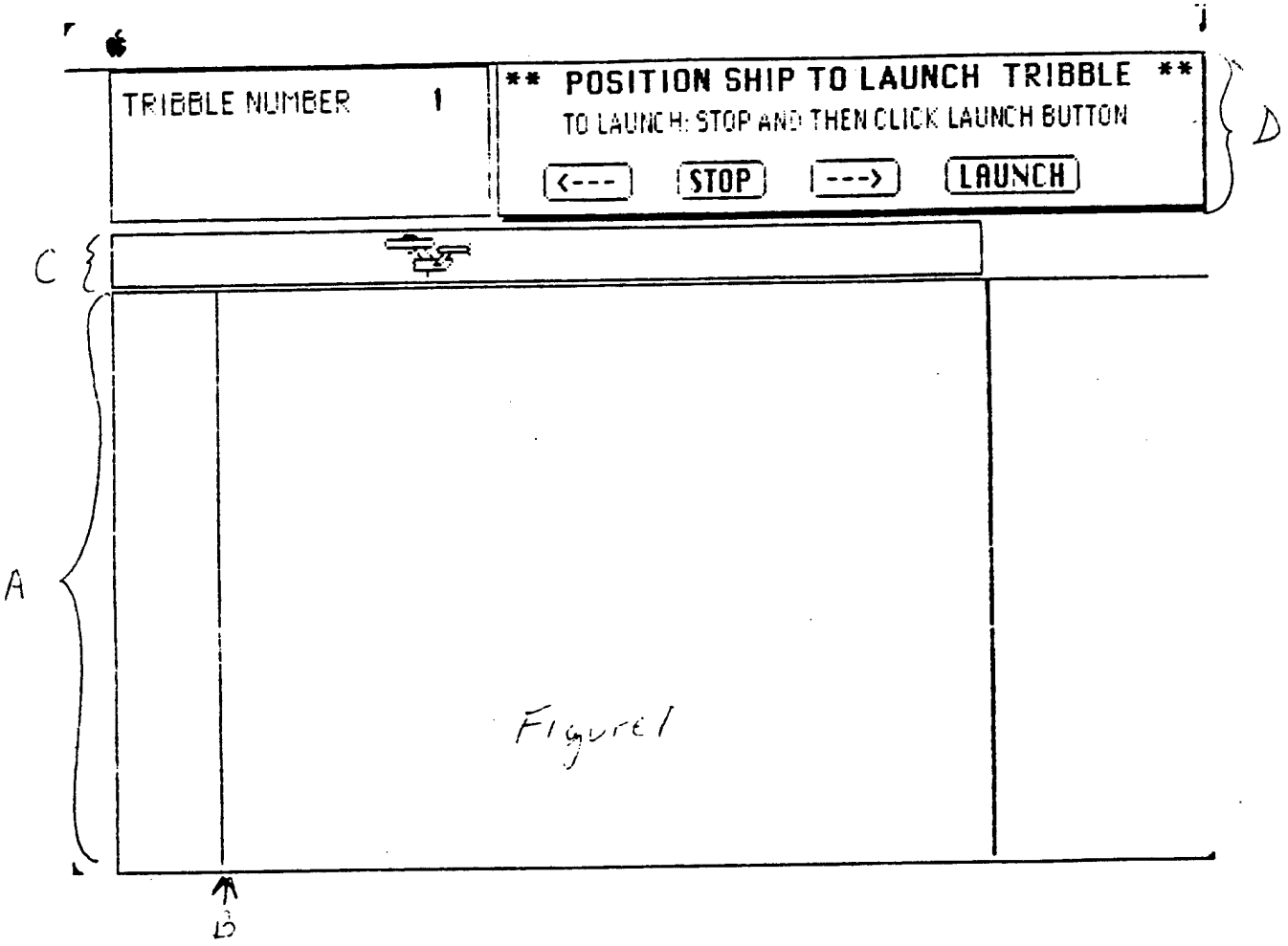


Figure 1

Fortunately, research has established that the distribution of moisture in the planet's soil is remarkably regular. The percentage of soil moisture on the planet's surface INCREASES uniformly from west (left side) to east (right side). Preliminary data indicate that tribbles can survive only when the percentage of moisture in the soil equals or exceeds a certain level. This level is indicated by the solid line drawn vertically across the planet's surface. (Labelled B in figure 1). Your job is to conduct a more thorough investigation of this tentative hypothesis that tribbles can survive only at moisture levels equal to or exceeding the level displayed by this line, by planting a colony of 600 tribbles east or west of the line and observing whether they live or die. Previous research has shown that if 450 (or more) of the tribbles survive, then the colonies will be self-sufficient. If less than 450 survive, the colony will die out. Thus, your task is to find where the line should be placed so that 450 or more tribbles survive.

Each time you are ready to plant a tribble colony, you will see your spaceship appear in the orbit-area above the planet (Labelled C in figure 1). At this time, you are ready to prepare to launch a tribble colony to the planet's surface. Position your spaceship launcher (the marker protruding from the bottom of the spaceship) exactly above the place you wish to plant tribbles. Positioning is conducted by the spaceship control panel in the upper right corner of your computer screen (Labelled D in figure 1). To activate the spaceship, you move the cursor/arrow (by moving the mouse) to the desired button, position the arrow within the button (as is shown on the figure), and then click the mouse <u>once</u> to activate the button. The arrow buttons designate the direction of spaceship movement. If you want to move the spaceship to the East, use the "---->" button. If you want to move the spaceship to the West, use the "<----" button. The "STOP" button stops the spaceship from moving in any direction. YOU MUST STOP THE SPACESHIP FROM MOVING IN ANY DIRECTION BEFORE YOU CAN EXECUTE ANY OTHER CONTROL PANEL COMMAND (SUCH AS MOVING IN THE OPPOSITE DIRECTION OR LAUNCHING A TRIBBLE COLONY).

****

Once the spaceship launcher is pointed directly above the place you wish to launch tribbles, begin the launching process by clicking the "LAUNCH" button on your control panel. Your spaceship's on-board computer controls the execution of the launching procedure.

Once the colony has been launched, a rectangle will appear on the screen under the spaceships' launcher. This rectangle represents the planting area at which you positioned your launcher to drop the tribbles. The rectangle shows where you intended the tribble to be planted. See Figure 2.

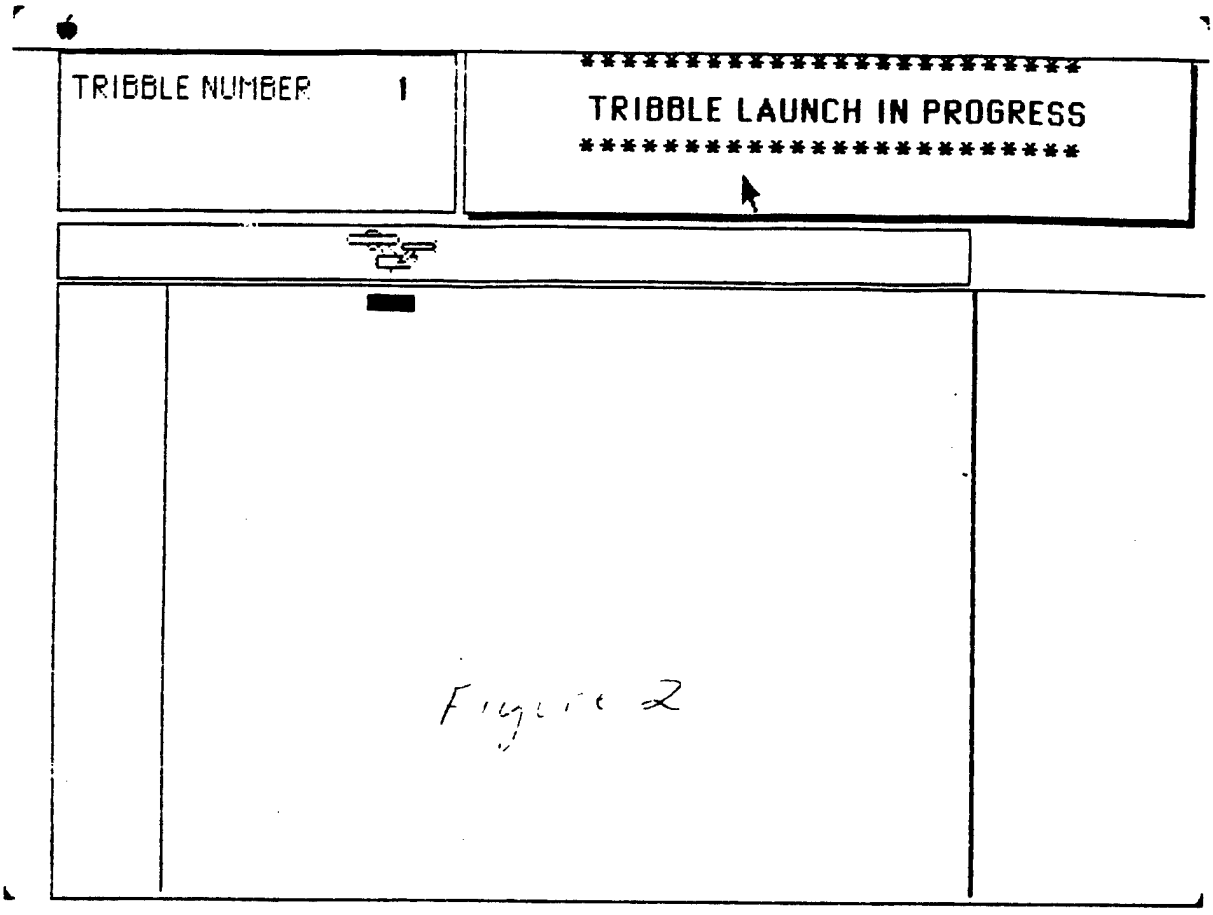


Figure 2

Why is the planting area represented by a rectangle instead of a single point? Because you are orbiting the planet from a distance of 500 miles, you are not able to drop the tribble colony at a precise location. Rather, tribbles land within 15 miles of the location at which you aim. Although the colony may land somewhat east, west, north, or south of the location you specify, only east-west error is of interest with respect to your moisture hypothesis, so this is the error you see represented on the screen by the rectangular box. Moving this rectangle east or west on each trial specifies a planting location. When the tribble colony is released, you will know that it has landed somewhere within the east-west range represented by the rectangle.

In order to determine where the actual moisture line lies, it is important to determine if the tribbles you planted lived or died at that planting location. So, every time you launch a tribble colony, one telemetry device is sent down with it. The telemetry device is located wherever the colony is planted. The device beams back to your ship information about how many tribbles in the colony lived or died. Previous

work with this device indicates that it is accurate within certain limits.

After you have given the command to launch a tribble colony, you will be asked to predict, based on the data you have available to you, whether the tribbles you are about to plant will live or die. Initially, you may feel you don't have enough information to warrant a reasonable prediction, but we'd like you to do your best. Remember that 450 or more of the 600 tribbles in the colony must live if the colony is to survive. See Figure 3

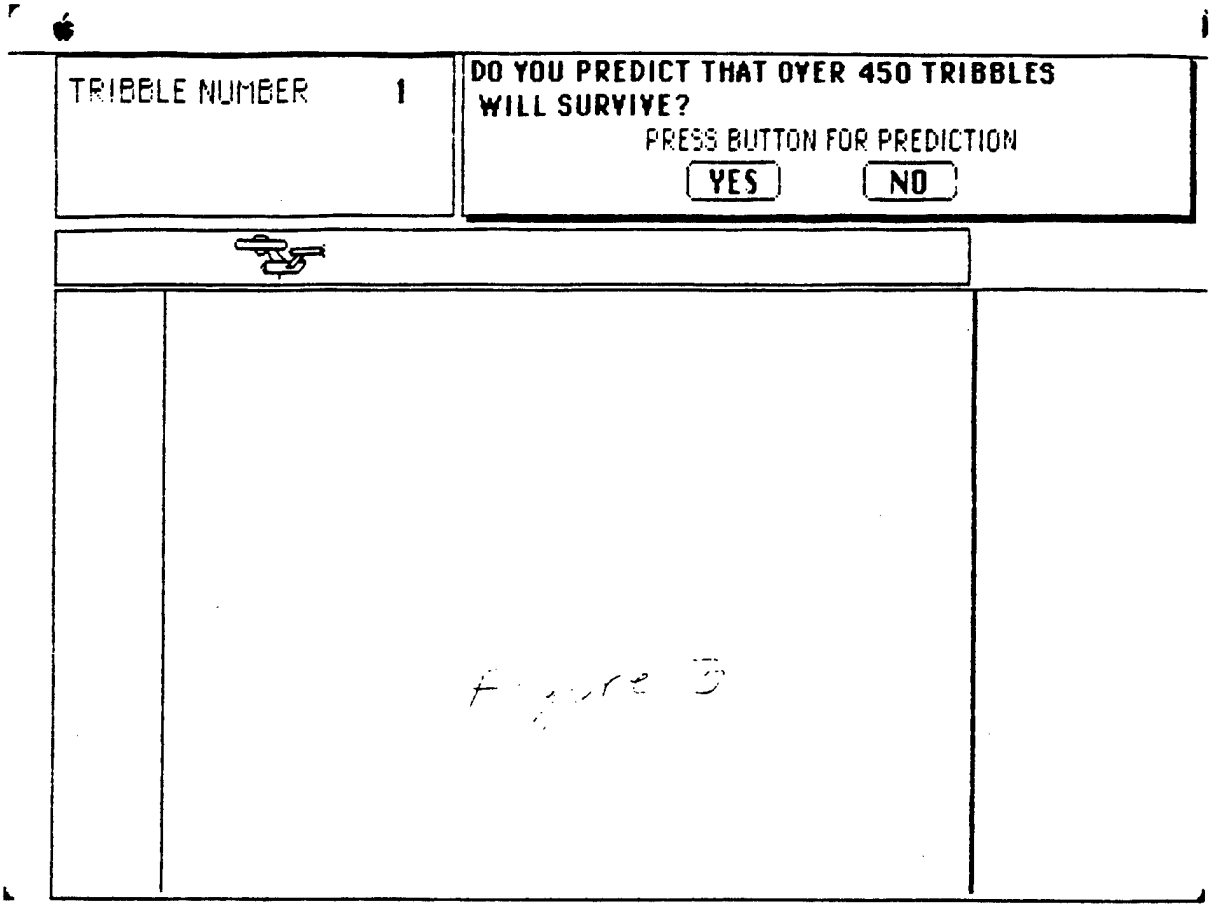


Figure 3

Indicate your prediction by moving the mouse and positioning the arrow on the response button desired. To make your prediction, you then click the button <u>once</u>.

On any one launch, some exact number of tribbles will survive, but you will not know this number. Instead you will get a range of possible numbers. At the right of the screen, the range of tribbles that lived is shown. (See figure 4.) You will know that the true number is somewhere within that range, but you won't know exactly where. Remember that at least 450 out of the 600 must survive for the colony to be able to survive.

You have resources available to plant a total of **8** tribble colonies.

After you have planted a tribble, determined how many have lived or died, and are preparing to launch your next tribble, a question will appear on the control panel of your computer screen which says "DO YOU WANT TO MOVE THE HYPOTHETICAL LINE?" (See Figure 4.)

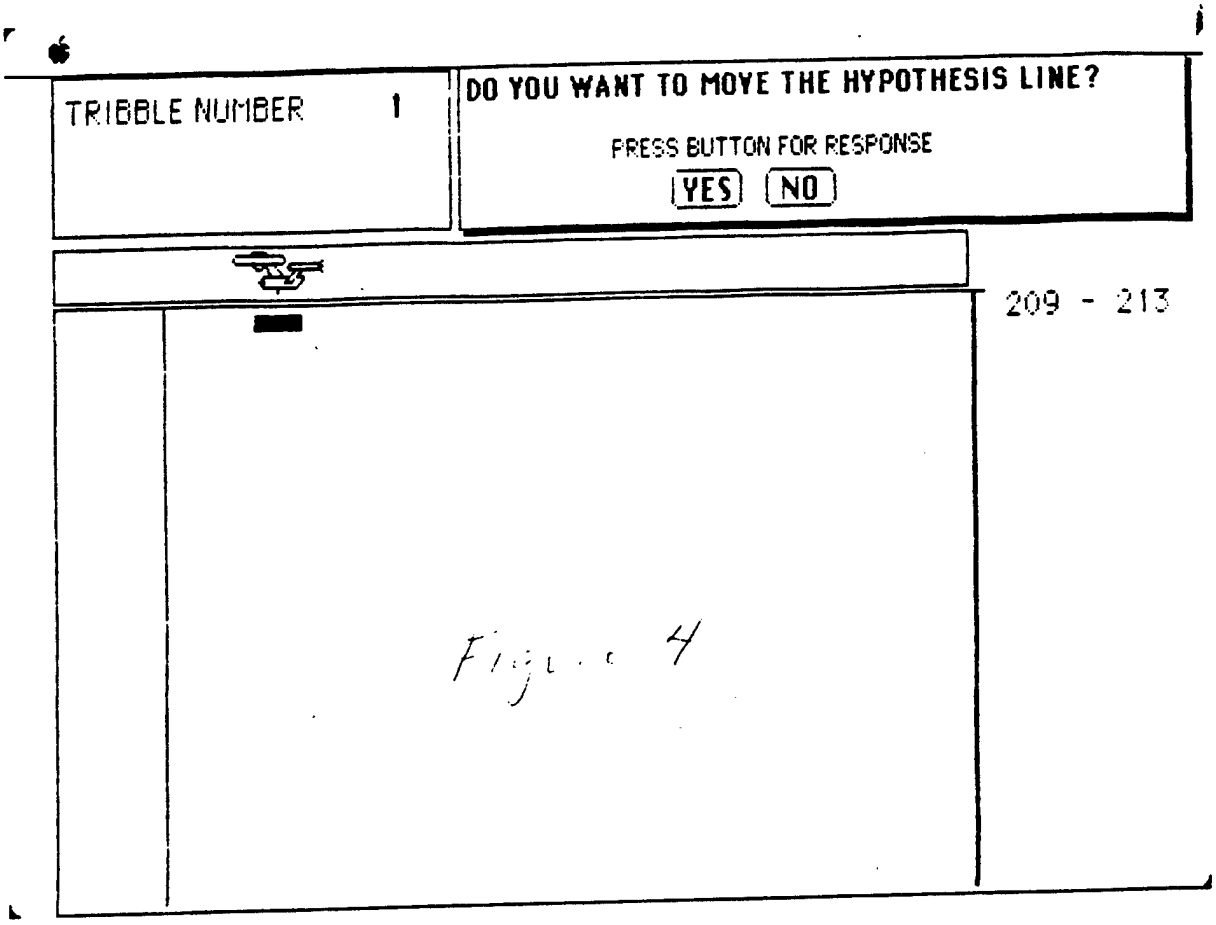


Figure 4

We are interested in finding out whether or not the data you've generated have led you to reject the location of the original critical moisture line in favor of a new location. Relocate this line when you feel your data indicate that the line's present position is incorrect and does not represent the actual critical moisture level. Again, respond by positioning the arrow over the desired response button (YES or NO) and clicking it <u>once</u>. Move the line as often as you wish. You should move it <u>every</u> time you feel its present position is wrong. Don't wait until the end, or try to save time by not moving it when you think it should be moved.

Even if you decide not to move the line, the computer will give you the opportunity to plant your next tribble colony. If you have decided to move the line, a line-moving control panel will appear, which works the same way as the spaceship-moving control panel. See Figure 5.

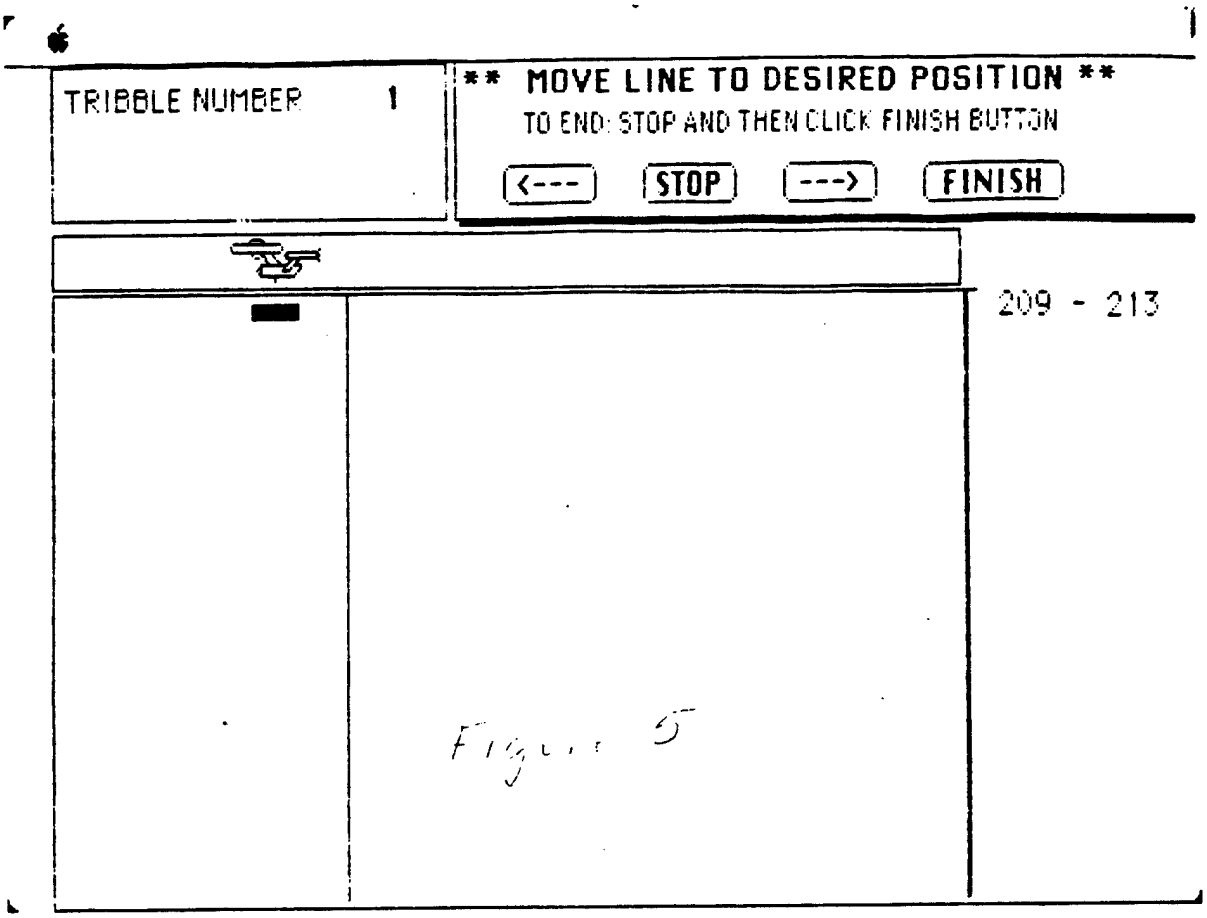


Figure 5

To move the line left or right, use the arrow buttons. Use the STOP button to stop movement in any direction.

When you've reached a point that represents your new working hypothesis about where the critical moisture line should be, click the "FINISH" button <u>once</u>. The line will be repositioned, and the computer will now give you your next tribble colony to plant.

This entire procedure will be followed for each tribble colony you plant until you've planted a total of **8** colonies.

To reduce possible confusion, there is a box in the upper left corner of your screen which keeps an account for you. Refer back to one of the figures and note the box. This information is updated each time you are preparing to launch a new tribble.

If you have any questions, please ask for clarification from the experimenter before the experiment begins.

Even if you decide not to move the line, the computer will give you the opportunity to plant your next tribble colony. If you have decided to move the line, a line-moving control panel will appear, which works the same way as the spaceship-moving control panel. See Figure 6.

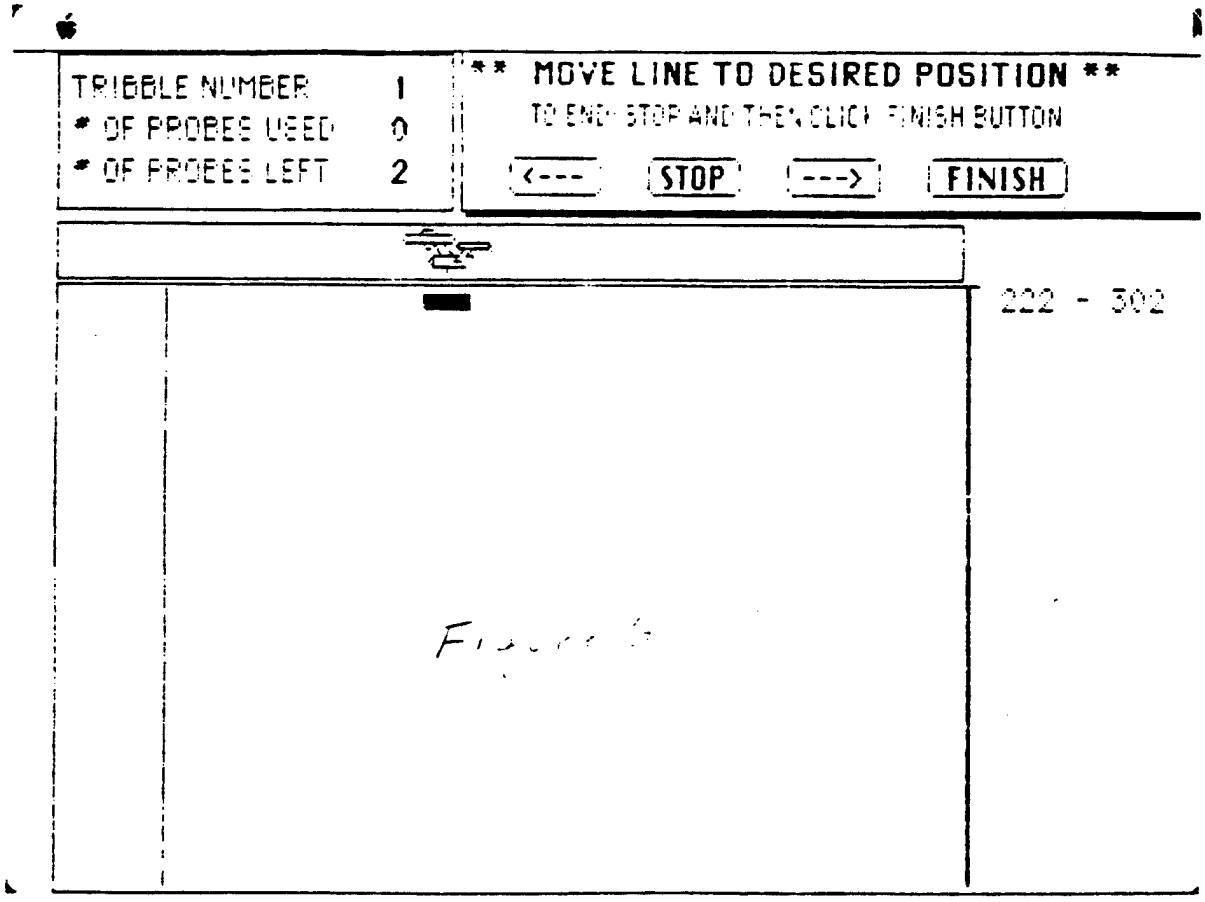


Figure 6

To move the line left or right, use the arrow buttons. Use the STOP button to stop movement in any direction.

When you've reached a point that represents your new working hypothesis about where the critical moisture line should be, click the "FINISH" button once. The line will be repositioned, and the computer will now give you your next tribble colony to plant.

This entire procedure will be followed for each tribble colony you plant until you've planted a total of **6** colonies.

To reduce possible confusion, there is a box in the upper left corner of your screen which keeps an account for you. Refer back to one of the figures and note the box. This information is updated each time you are preparing to launch a new tribble.

If you have any questions, please ask for clarification from the experimenter before the experiment begins.

If such a malfunction occurred, then the feedback received on the screen would bear no relationship to the actual number of tribbles that may have lived. In order to assess such a malfunction, subjects in the high SF error conditions were able to conduct two probe checks. A probe check could tell a subject whether the feedback received was accurate (in which case the number of tribbles that lived was within the range displayed on the screen) or whether the feedback was inaccurate (the number of tribbles that survived could be anywhere between 0 and 600). A probe check could be conducted after any launch, but subjects were allowed only two probe checks in the first eight launches and one additional probe check in the last four launches.

Results

Because the design of the two experiments is similar, it will be convenient to describe the results together. Several different dependent variables will be discussed. The reader should bear in mind that Experiment 11 is a 3 X 3 factorial in which ME error and SF error were manipulated on the input side, whereas Experiment 12 is a 2 X 2 factorial in which the two kinds of error were in the feedback.

Accuracy of Final Hypothesis. In each experiment, the distance between a subject's final moisture level hypothesis line and the actual moisture level line provides a measure of overall performance. For Experiment 11, this distance is plotted as a function of condition in Figure 4-4. Note that greater distance is equivalent to poorer performance. Figure 4-5 displays the equivalent mean distance for subjects in Experiment 12 as a function of experimental condition. Here also there are apparent effects due to SF error and to ME error.

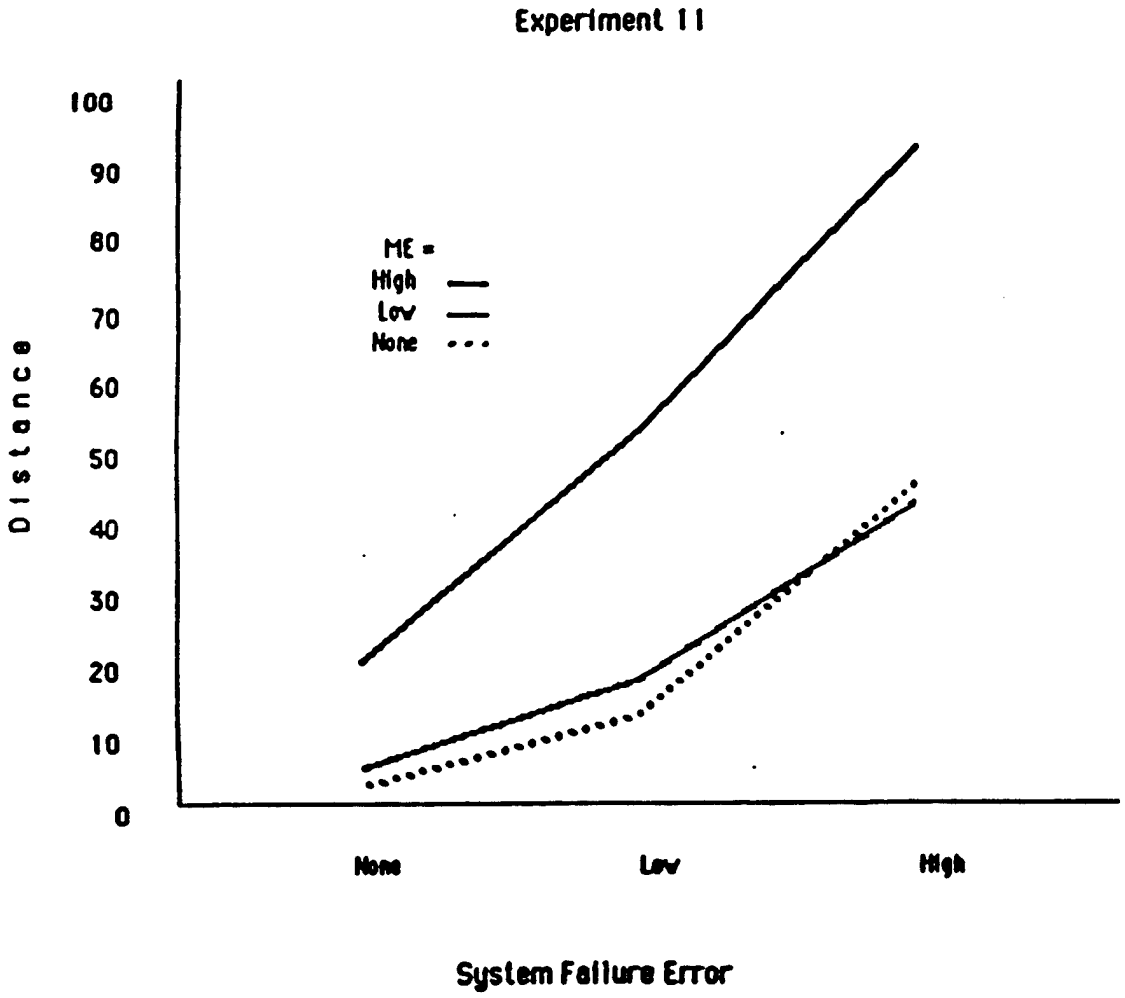


Figure 4 - 4 Mean absolute deviation of the subjects' final hypotheses from the true critical line.

Experiment 12

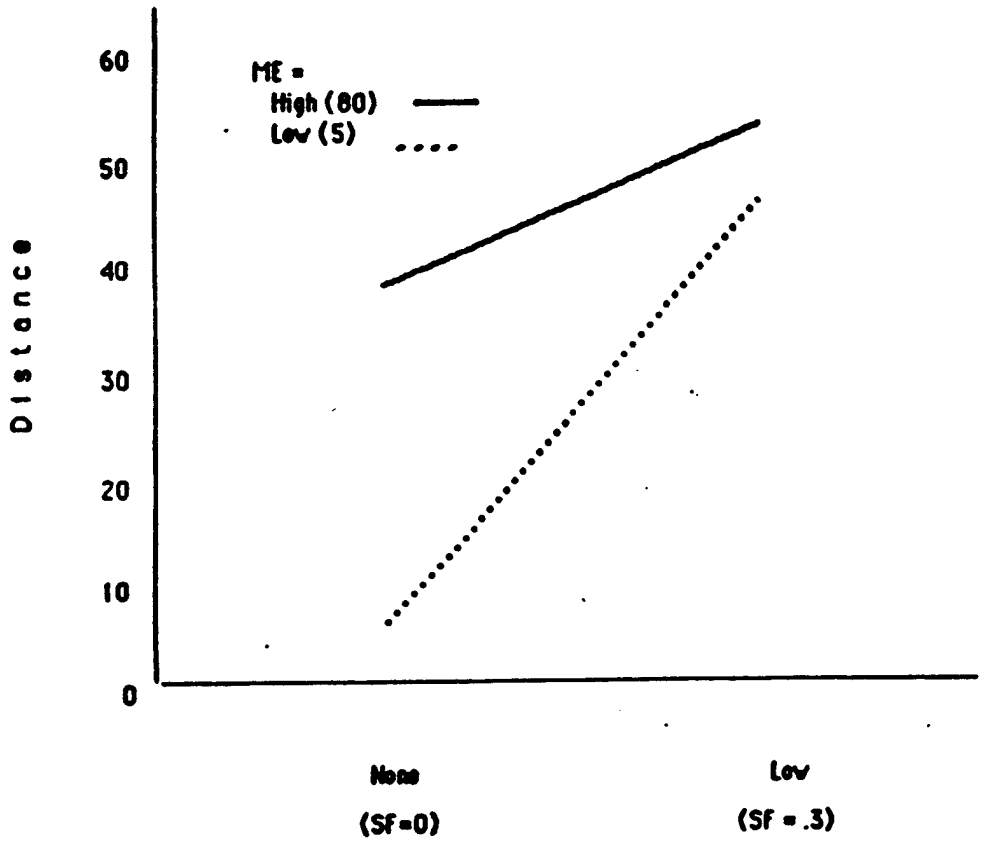


Figure 4 - 5. Mean absolute deviation of the subjects' final hypotheses from the true critical line.

Inspection of the data for both experiments revealed substantial heterogeneity of variance across experimental conditions and marked departure from normality as well. The data were accordingly analyzed using Kruskal-Wallis One-Way ANOVA by Ranks separately for each main effect. In Experiment 11, both main effects were significant; $H(2) = 18.70$, $p < .001$ for SF error, and $H(2) = 7.92$, $p < .02$ for ME error. In Experiment 12, the effect of ME was significant; $H(1) = 14.16$, $p < .001$, but the effect of SF error was not; $H(1) = 1.38$, $p < .30$.

Experiment 11 thus replicates the results of Kern's study insofar as ME error effects were found on the input side in both Kern's study and in ours, though we found a larger difference due to ME error on the input side than had Kern. Experiment 12, however, failed to replicate Kern's finding of a significant SF error effect on the feedback side, although it should be noted that a trend in the correct direction was found.

Because experimental condition in both experiments affected the amount of variance in the distance measures, individual subjects' data were plotted. Figure 4-6 presents this data for Experiment 11 and Figure 4-7 presents this data for Experiment 12. In both cases, distances plotted are absolute values, that is, whether the subject's final hypothesis line was to the right or to the left of the actual line was not taken into account.

Inspection of the figures reveals a pattern common to both experiments. In both cases, it is clear that a small number of subjects are contributing to the mean differences. SF error at high levels appears to have greatly disrupted a small number of subjects in both experiments, while leaving a majority of subjects in both experiments relatively

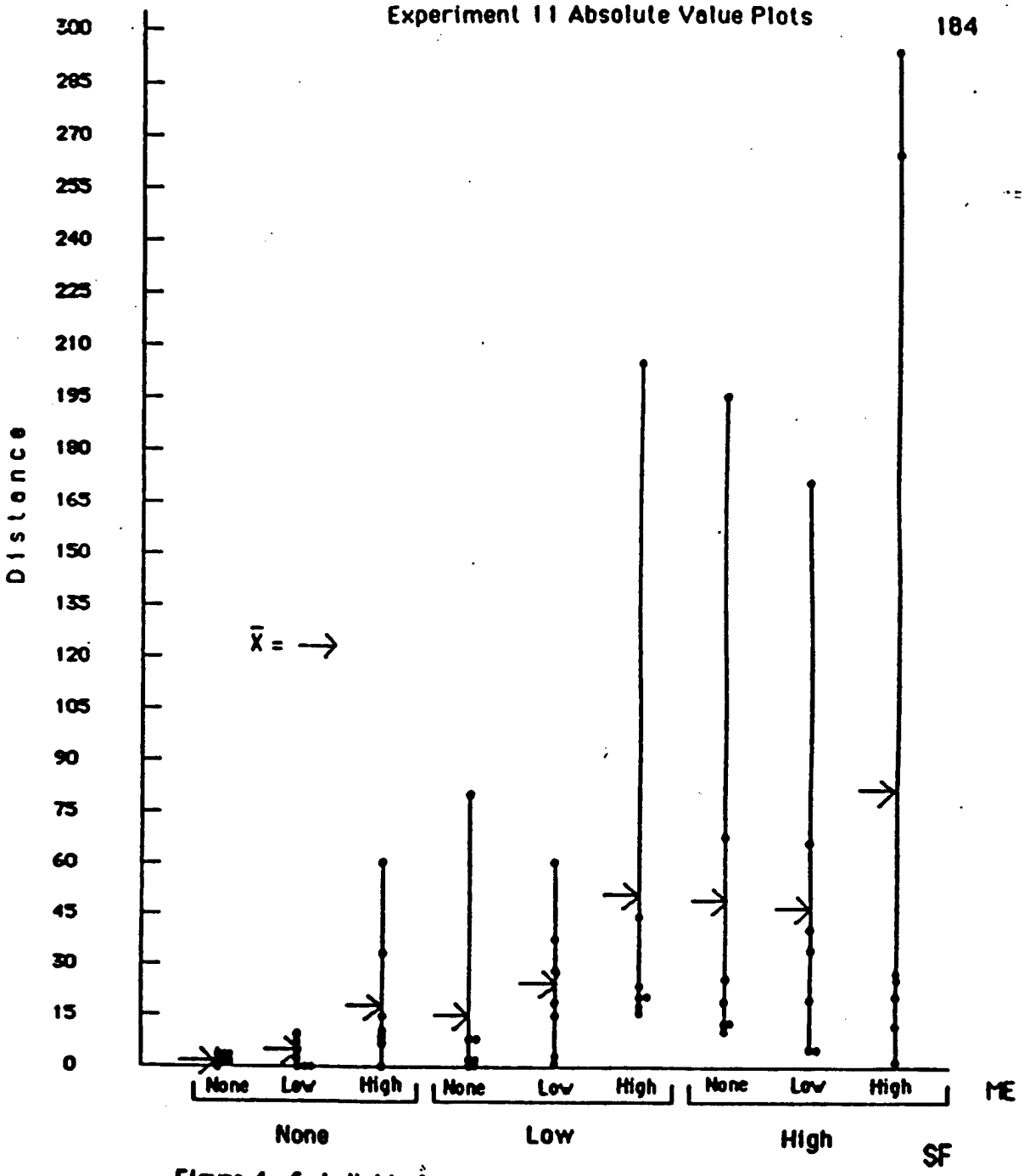


Figure 4 - 6. Individual subjects' accuracy scores (absolute deviations of their final hypothesis from the true critical line for experiment 11. conditions are:

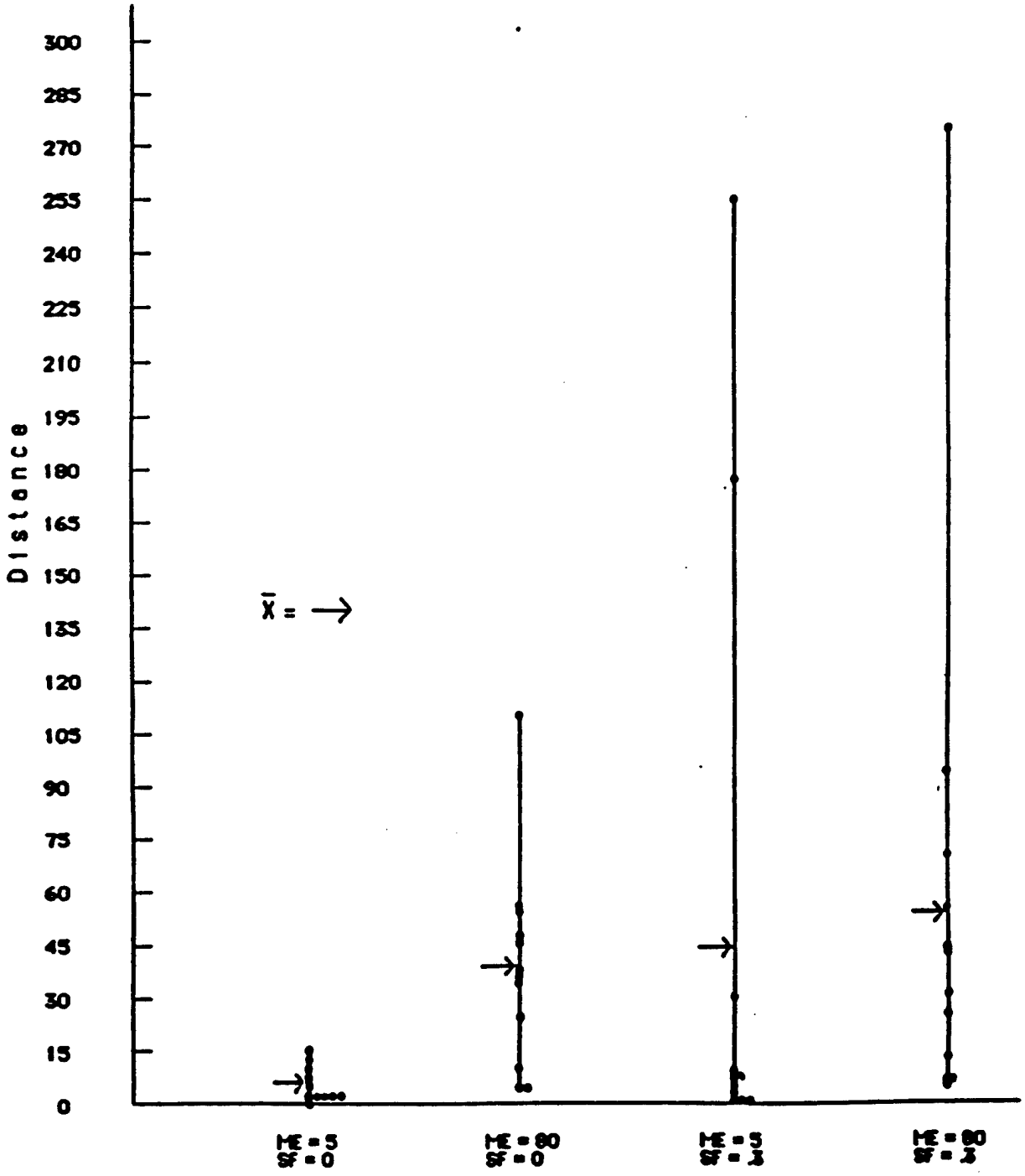


Figure 4 - 6. Individual subjects' accuracy scores (absolute deviations of their final hypothesis from the true critical line for experiment 12.

unaffected. High levels of ME error produced a similar pattern, though fewer subjects were involved, and even those did not generally manifest the extremes of disruption caused by SF error. Figure 4 - 7 shows that Kern's data manifested very a similar pattern (recall that she used a 2 X 2 design with two levels of ME error and two levels of SF error).

Figures 4 - 5, 4 - 6, and 4-7 also suggest that the effects of ME error are more consistent than those of SF error. Whereas SF error caused a small proportion of subjects to do very poorly, increasing levels of ME error appeared to cause a large number of subjects to perform slightly less well.

Subjects' hypothesis testing strategies. Does error influence the way that subjects attacked the problem? In both experiments 11 and 12, the pattern of tribble plantings was a variable of key interest. To assess the subjects' patterns of plantings we chose to define a quantitative index which reflects our conception of a good strategy of attack in this particular task.

A good strategy, given the instructions, would have been something like a systematic eastward progression of planting, perhaps with a successive halving of the distance between the last planting and the easternmost edge of the surface being surveyed. Figure 4-8 presents an example of a good strategy. An extremely poor strategy would have been to plant tribbles randomly (see Figure 4-9 for an example of a poor strategy). There are strategies that would have been even worse than random planting, e. g., progressive westward planting, but that would have been completely out of character with the nature of the task. Therefore for scaling purposes we defined random planting as essentially the worst plausible strategy,

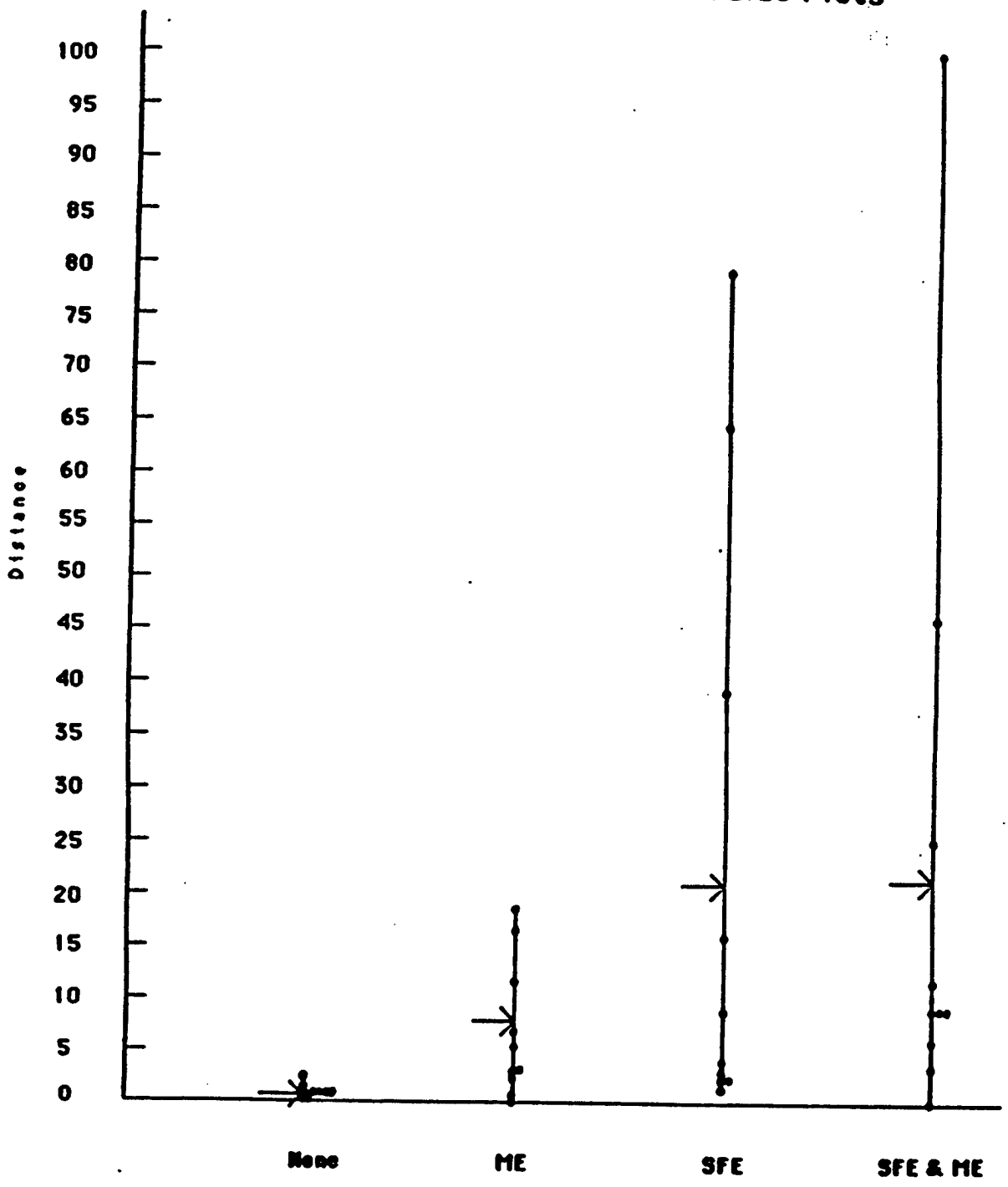


Figure 4 - 7. Individual subjects' accuracy scores (absolute deviations of their final hypothesis from the true critical line for Kern's data.

and finding the line immediately and planting on it on all 12 trials as the best possible strategy. A numerical value that varies in the appropriate way can be obtained from the differences between successive plantings. Such a difference was used as the basis for a "goodness of strategy index", with higher numbers indicating poorer strategies.

The first step in determining the goodness of strategy index was computing the average absolute value of the differences between successive plantings. This is conceptually a reasonable measure, but several subjects, after they had quickly and successfully narrowed down their search and located the line quite accurately, apparently used one or two observations as "checks", and went far back to the west to plant a tribble or a colony of tribbles. This behavior seemed eminently reasonable, but it drastically inflated the strategy score. Hence, we operationally defined "checks", and removed their influence from the index. A check was defined as an observation that met all three of the following criteria: 1) the planting was done after the subject's hypothesis about the critical line had stabilized, i. e., after the last move of $\leq$10 pixels, 2) the difference from the last planting was large, i. e., $\geq$ 50 pixels, and 3) the subject predicted that the tribble or the colony of Tribbles would not survive. The strategy index was 20.6 for Figure 4-8 and 136.5 for Figure 4-9.

In Experiment 11, both ME and SF error had significant effects on the strategy index; $F(2,54) = 3.75$, $p < .05$; $F(2,54) = 4.23$, $p < .05$, respectively. The ME X SF interaction was not significant; $F(4,54) = 2.12$, $.10 > p > .05$. Figure 4 -10 shows that the effect is essentially due to the High ME-High SF condition, with a pattern index almost twice as high in that condition

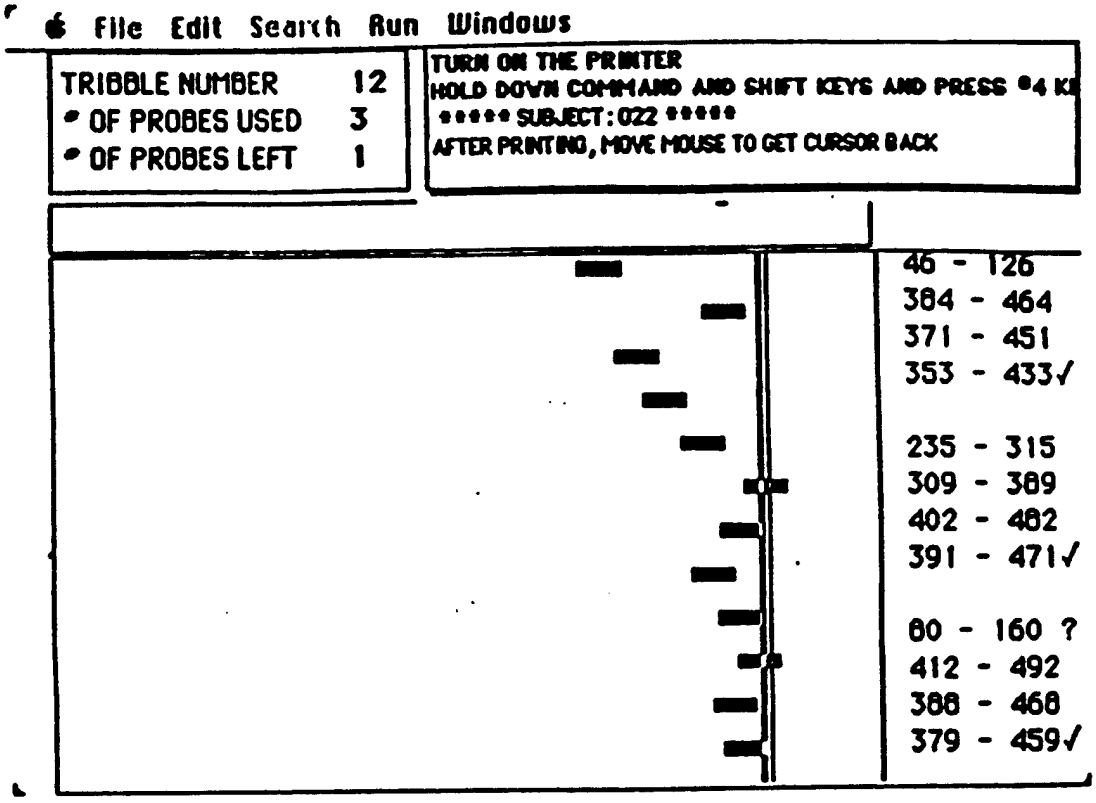


Figure 4 - 8. An example of a final screen representing a good pattern of tribble launches.

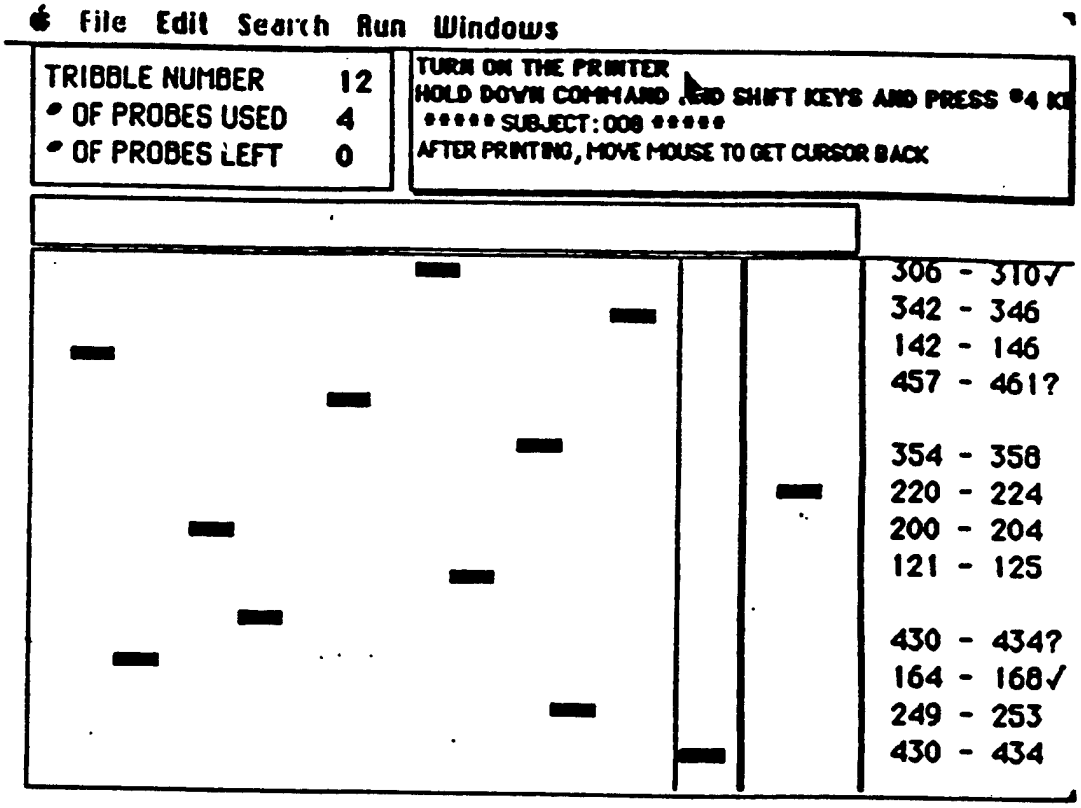
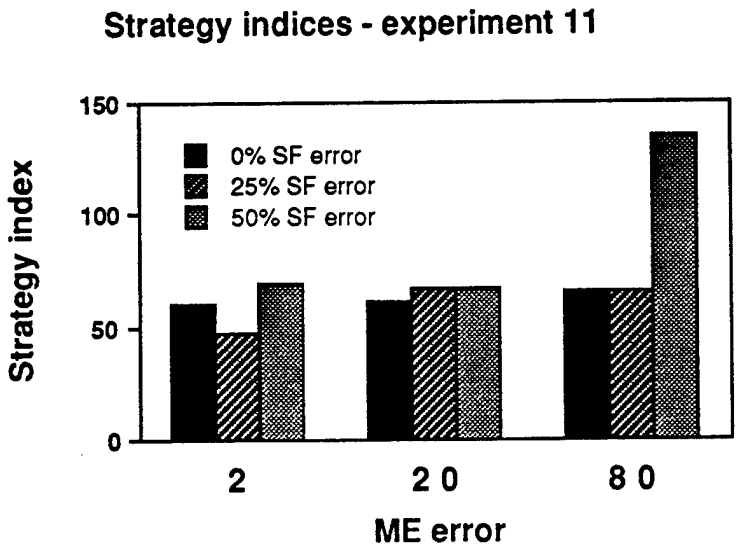


Figure 4 - 9. An example of a final screen representing a poor pattern of tribble launches.

as in any other, and with relatively similar means in the other eight conditions.



In Experiment 12, there was no effect of ME error on the pattern index ($F<1$), nor was there evidence of an interaction between ME and SF ($F<1$). There was, however, a significant effect of SF error; $F(1,44)= 8.79$, $p<.01$, with poorer strategies in the conditions with SF error present.

Hypothesis revisions. Kern found that subjects were much less likely to move their hypothesized moisture line in the presence of SF error than in its absence (ME error had no effect). The effect she found was quite large; subjects given SF error in the feedback revised their hypothesis line less than half as often as subjects given no SF error. We found no comparable effects; in neither experiment was there a significant difference in number of hypothesis revisions as a function of SF error. In fact, the only significant result in our data was that subjects given high ME in

experiment 12 were more likely to revise their hypotheses; $F(1, 44) = 4.83$, $p < .03$. The effect was slight, however, the mean number of revisions being 6.4 for low ME and 7.8 for high ME, representing less than 10% of the total variance in the number of hypothesis revisions.

Probe checks. One of the dependent variables that we expected to be sensitive to error effects was the number of times subjects checked to see if the telemetry had malfunctioned. Kern found that subjects were more likely to query a probe which disconfirmed their expectations than a probe which confirmed their expectations, but her effects were only marginally significant. We did not replicate her findings: in neither experiment was there a significant effect of error on the frequency of probe checks.

Replications. In the analysis of the 2-4-6 task we were concerned with whether SF error led subjects to attempt to assess the system unreliability b y systematically replicating trials (triples). A similar analysis could not be easily accomplished in the present study, because it was not clear whether a subject was trying to launch a tribble from the same position, since they were visually locating the final ship position. Also, since the number of observations was strictly limited in the artificial universe task, a replication was likely much more costly to the subjects. Nevertheless, inspection of the data suggests strongly that there was very little attempt by subjects in any of the groups to use their tribble resources to check system reliability.

The relation between strategy and accuracy. There are a large number of possible relations among dependent variables, but the most important possibility is between the strategy and the accuracy scores. In experiment

11, the overall correlation between these variables across groups was .63. This is a remarkably high correlation, given the many influences on both of the underlying constructs, and given the relative crudity of the indices of those constructs. The corresponding value for experiment 12 was positive, but lower, .26. In experiment 12, the correlations between strategy and accuracy were higher within the four separate groups than the overall .26, in two groups considerably so. These values cannot be legitimately tested for statistical significance, for while the Pearson r can be used as a statistic descriptive of the linear component of a relationship under almost any circumstances, the available statistical tests require bivariate normalcy, a condition not met in our data.

Discussion

Though not exact replications of Kern's study, experiments 11 and 12 are suggestive in their implications for the generalizability of her results. We found, as did Kern, that SF error can have major implications for human performance in complex systems. While not a major focus of her study, we found, as she did, that SF error can lead to major performance deficits in a small number of subjects while affecting most subjects very little. ME error was more consistent in this regard: in both Kern's study and ours, presence of ME error affected most subjects to a small extent. It seems clear that a full understanding of the effects of SF error will need to pay very close attention to the specifics of individual differences in performance. By contrast, a full understanding of ME error effects may be possible using the nomothetic approach common to most earlier studies of the effect of error on performance. Overall accuracy aside, Kern's study found large and consistent changes in the strategies used by subjects in the presence of SF error. We

found less strong effects. It is not clear why this difference exists, but it is plausible to suggest that the need to unconfound locus of error and type of error led to a more complex task environment for our subjects. This seems especially likely in experiment 12; in general, we were not convinced that all of our subjects were fully cognizant of the nature of the feedback information, in spite of the extensive instructions. For this reason, we are not as confident about the outcome of experiment 12 as we are of the outcome of experiment 11.

A potentially important finding was the relation between the error manipulations, the strategy indices and the accuracy scores. While much different experimental designs would be needed to tease out the nature of the effects, note that in both experiments 11 and 12 there were significant effects of error on strategy, and in both there were positive correlations between strategy and accuracy. The people who did just especially poorly on the task tended to be those with high levels of SF error, and with high strategy (i. e., poor) strategy indices. Unfortunately, this task, unlike the MCPL task, allows no easy way to separate the purely informational effects of information degradation from effects that go beyond the information lost.

In general, the two studies do suggest that the differential effects of locus and type of error found in the first 10 experiments are generalizable to more complex environments. In spite of the complexity of the outcome, it seems true that SF error, in particular, has unique implications for performance. For this reason, we believe that experiments 11 and 12 constitute strong evidence in favor of close attention to this variable on the part of designers of complex real-world interfaces between people and fallible sources of information.

Part 5

DISCUSSION AND IMPLICATIONS

The primary contribution of the research described in this report is the elucidation of the concept of system failure error, and the start of a serious, systematic attempt to determine the conditions under which it occurs, and the effects that it may have. The idea was described vividly by Seymour Hersh, in his controversial book on the downing of the KAL 007:

> The fact is that many pilots simply do not rely on ground - mapping radar because ... they don't believe it will tell them anything they need to know, and when it does depict conflicting data, they frequently choose to believe that it is malfunctioning.

Such error is associated with man-made artifacts, with technology. It is associated with computers, perhaps especially vividly in those occasional newspaper articles that report some hapless soul who has just gotten a telephone bill for several billion dollars. The response described by Hersh amounted to ignoring the output of a man-made system that operators had too often seen fail. When a radar goes awry, it may not just involve a ±1% or ±2% error, it may involve a system output that is fundamentally unrelated to the process that the system is supposed to be representing. The effects of this form of error on the person who must deal with the system are virtually unstudied.

In this report we contrast SF error with normally distributed, or "measurement error." It is the sort of error that we associate with measurements made, for example, with a yardstick, or a tape measure. If we were to ask a number of people to measure one person's height to the

nearest tenth of an inch with a tape measure, we would get a distribution of measurements which would be reasonably normal in shape. According to classical measurement theory, or classical psychometric theory, the mean of an infinite number of such measurements would be that person's true height, and the variation in the measurements would be an index of the reliability of the measuring process.

SF error is quite different. It is the form of error that characterizes many technological systems. It is the form of error that Hersh referred to when he characterized the system as "malfunctioning." With respect to that aspect of the man-machine interface by which the human gains information about the world from various system indicators, little hard knowledge exists about the effect of system errors on the user, even in redundant systems where such error would mean conflicting output. So far as we know, little is known beyond the sort of anecdotal evidence cited by Hersh, part of which is quoted above.

The second major aspect of error that underlies the research effort under this contract is the "location" of the error in the sequence of operations between the human and the information system. We define the information on which some inference or prediction is based as input (note that we are referring to input from the system to the operator). Information about the correctness of the inference or prediction is defined as feedback (from the system to the person). Hence, two loci of error are possible: we may have input error or feedback error. The 2x2 error taxonomy just described is by no means exhaustive, but it suggests at least four sorts of error that may have important consequences for behavior.

Generalizability from the laboratory to the world. This is often a major

concern of experimental psychologists. With respect to the present research, however, we believe that any effects demonstrated in the laboratory are likely to be exacerbated in real operational contexts. This is because in the experimental situation all subjects had to do was to make an inference or prediction about the system, based on evidence only from the system, and had as much time as they desired to do so. In operational contexts, the person will have to make highly consequential decisions based on the inferences or predictions, will be operating under time pressure, and may have a variety of sources of conflicting information. All of these will, we believe, exacerbate the effects of error, especially SF error.

What are those effects?

SUMMARY OF FINDINGS

1) Performance is **not** disrupted by ME error to a degree greater than performance would be expected to be degraded by the loss of information entailed by the introduction of error. This does not mean that ME error has no effects; it means that subjects basically ignore such error and very often that is the optimal thing to do. There are dependent variables that show ME effects. For instance, Markowitz's unpublished research showed that subjects do not show regressiveness in their predictions appropriate to the unreliability inherent in the concept of ME error, and York, Doherty & Kamouri (1987) showed that subjects believed that ME error was quite disruptive.

2) Performance **can** be badly disrupted by SF error, to a greater degree

than would be expected by the loss of information entailed by the introduction of error. This was shown unequivocally in the MCPL paradigm, which was the only paradigm used that allowed us to equate statistically the error types and that had a measure of optimal performance. In the MCPL work we could show that the SF effects were in fact greater than the effects of the information degradation entailed in the introduction of error.

3) The effects of SF error are highly variable: some people are much more seriously affected than others. The magnitudes of the effects of SF error on the individual differences were a surprise.

4) Subjects' beliefs about the effects that error has on them may differ from the actual error effects. This finding warrants careful investigation. The operator's beliefs about the systems may turn out to be the key as to whether the system is used as designed, and to the huge performance deficits shown in some subjects. Some evidence not consistent with this is the failure to find effects of informing the subjects of the presence of error in Wason's rule discovery task (experiment 8).

5) Many subjects display agitation when faced with SF error. This was an anecdotal observation made by the experimenters as subjects tried to solve Wason's rule discovery task and the artificial universe task. It is a phenomenon we noticed many years ago in perceptual tasks, as well (Doherty & Keeley, 1972; Keeley & Doherty, 1971a,b). People like to be right, even when absolutely no payoffs are involved. They get upset when they cannot trust their senses to tell them the truth about the world, as in the perceptual tasks. They also get upset when they cannot trust technology to tell them the truth about the world, as in the tasks described here. This may also be related to the speculation on p. 68 that the big jump

in the "psychological cost" of being wrong occurs at the difference between a direct hit and a small error, whereas the almost universally used loss functions are power functions of the magnitude of the subject's error.

6) Subjects tend strongly to attempt confirmatory tests of the hypothesis under test, rather than to test alternatives or to attempt disconfirmatory tests. This is not a motivationally mediated phenomenon. The bias to confirm is, of course, a well-known phenomenon (Mynatt, Doherty & Tweney, 1977;1978; Snyder & Swann, 1978), replicated in the course of this research.

7) The availability of highly diagnostic information can attenuate the otherwise robust bias to confirm. This phenomenon has also been previously reported, but it is not well-established (Skov & Sherman, 1986; Slowiaczek & Sherman, 1987; Trope & Bassok, 1982;1983). Our reading of this literature is that there are both effects which are confirmatory in nature, but that diagnosticity - the degree to which the data ought to influence one's belief in some hypothesis relative to the alternative(s) - also has an influence that may or may not override the tendency to confirm. What we now believe to be a source of the tendency to confirm will be discussed below.

8) Subjects like big numbers. This is a rather bizarre bias that occurred in the pseudodiagnosticity research outlined in Part 3. There was a marked tendency for subjects to select large percentages, regardless of the diagnostic impact of those values. This is a wholly dysfunctional cognitive bias.

9) Subjects did not check the reliability of the data very often. While there were, in the 2-4-6 research, significantly more repeated

observations in the SF error conditions than in the no error conditions, our judgment is that the subjects made far fewer such test-retest reliability checks than they should have. This was also true in the artificial universe experiments, but in those experiments the subjects were limited in the number of tribbles they could plant. There was no limitation in the 2-4-6 task on the number of triples to be tested before rule announcement, and the subjects could, with minimum effort, retest triples to assess not only the possibility of error on any given trial, but also the seriousness of the problem created by the presence of the error.

METHODOLOGICAL RECOMMENDATIONS

Future investigations should ideally include protocol analysis as well as a variety of behavioral indicators. Hindsight leaves us wishing that we had selected a small number of subjects in many of the studies, and taken "think-aloud protocols" on them. Such protocols would have provided, at the least, hypotheses about the cognitive processes employed by subjects in dealing with these tasks. Perhaps protocol data would have provided strong converging operations for some of our conclusions regarding the effects of SF error.

The major new findings of the research conducted under this contract deal with the effects of SF error. In one study (Experiment 11), we found drastic effects, but with relatively few subjects. How would we proceed in trying to make this knowledge useful? One possibility would be to adopt an individual differences approach, and to look for personality correlates of susceptibility to SF error. We do not think that this would be a fruitful

way to go, though there are those who would argue for using the concept of "cognitive styles". A second possibility would be to try to discover the environmental conditions which differed between the subjects who did badly and those who did well. In the artificial universe studies what happens to each subject on any given trial differs as a function of chance factors, and as a result of the subject's behavior on previous trials. We do not have sufficient data from each subject in Experiment 11 to home in on one explanation and also rule out others.

We have been poring over error data and thinking about the effects of error in cognitive tasks for many years, and we think that the most productive next step would be to attempt to assess operators' beliefs about specific interfaces after a controlled experience with those interfaces, experiences that are analogous to our ME and SF manipulations, especially the latter. If Hersh's informants were correct, and if our subjects and tasks are representative, then SF error may have disastrous consequences-and operators certainly ought to be able to report when they lose faith in some system.

SPECULATION ABOUT THE COGNITIVE SOURCES OF ERROR EFFECTS

We believe that the effects of error are due in part to a fundamental limitation on how many things a human being can think about at one time. The assertion that we can think of but one thing at a time is, we think, phenomenologically obvious and consistent with the data, but one that we do not see explicitly in the psychological literature. How many simultaneous objects of thought, of "internal attention", can there be? The

answer seems self-evident, one.

"Thing" may refer to an object, or to an attribute, or to a relation between objects or attributes. One can think of the relation between a symptom and a disease, but not, simultaneously, of the relation between a symptom and two diseases: or at least not without unusual effort. "Simultaneous" must also be carefully defined. Clearly, people can switch internal attention more or less rapidly from one focus of that attention to another, although we believe that this happens far less rapidly than is the case with external attention. We do not claim that it is impossible to think about two things, only that it is impossible to think about two things at the same time. If this is true, it follows immediately that an attempt to, say, think about the relation of a symptom to two diseases, **must** involve the switching of internal attention. We hypothesize that such switching is difficult in proportion to the complexity of the relations between the alternatives and the information, and that all other things being equal, people avoid difficult things. It is easier to think about only one alternative and ignore the other one. The reader may have had the experience of trying to decide which of two scientific theories a set of data best fits. It's hard to do. Almost invariably cognitive "aids", such as written lists or notes, are necessary. It is simply too difficult to keep all of the relevant relationships in the head.

The above paragraphs borrow heavily from Doherty & Mynatt (1986), a paper that dealt with the pseudodiagnosticity phenomenon. It is an attempt to explain why people select the inappropriate probability values in that task. When we first began thinking about this explanation, Tweney recognized its relevance to the 2-4-6 task, the optimal approach to which

requires subjects to entertain and test multiple hypotheses, or to entertain both the hypothesis that "numbers increasing by 2" is the correct hypothesis, and also that the same hypothesis is the **in**correct hypothesis. Multiple hypothesis testing and falsification are cognitive activities that some people claim they engage in routinely, but which we very rarely see in the laboratory even when we instruct people to engage in them (Mynatt, Doherty & Tweney, 1977).

What is the relevance to error effects? We think that it is very hard for subjects to entertain simultaneously hypotheses about the state of the world implicated by the data **and** hypotheses about the data, or about the data source. Once our working memory is taken up by hypotheses about the data or the data source, we can no longer concentrate on drawing an inference about the present state of some aspect of the world, or predicting some future aspect of some state of the world. But it is the state of the world that is of interest, not the data. Hence, if something has to be subordinated, it will be the data - we will stop attending to the source of the data that competes for attention, and rely on other sources that may be less valid indicators, but that we can rely on to give us reliable information.

References

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudo-diagnosticity. Journal of Personality and Social Psychology, **45**, 1185-1195.

Brehmer, B. (1970). Inference behavior in a situation where the cues are not reliably perceived. Organizational Behavior and Human Performance, **5**, 330-347.

Brehmer, B. (1980). In one word: Not from experience. Acta Psychoologica, **45**, 22-241.

Brunswik, E. (1956). Perception and the representative design of psychological experiments. Berkeley, CA: University of California Press.

Castellan, N. J. (197). Decision making with multiple probabilistic cues. In N. J. Castellan, D. P. Pisoni & G. R. Potts (Eds.). Cognitive theory,Vol. 2. Hillsdale, NJ: Erlbaum.

Doherty, M. E., & Balzer, W. K. (In press). Cognitive feedback. In B. Brehmer & C. R. B. Joyce (Eds.). Human judgment: The social judgment theory approach. Amsterdam, North-Holland.

Doherty, M. E., & Keeley, S. M. (1972). On the identification of repeatedly presented, brief visual stimuli. Psychological Bulletin, **78**, 142-154.

Doherty, M. E., & Mynatt, C. R. The magic number one. (1987). In D. R. Moates & R. Butrick (Eds.) Proceedings of the Ohio University Interdisciplinary Inference Conference. 221-230.

Doherty, M. E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. Acta Psychologica, **43,** 111-121.

Doherty, M. E., Schiavo, M., Mynatt, C. R., & Tweney, R. D. (1981). The influence of feedback and diagnostic data on pseudodiagnosticity. Bulletin of the Psychonomic Society . **18**, 191 -194.

Doherty, M.E., Rothstein, H., & Schipper, L.M. (1983). The influence of two error types on predictions. Paper presented at the annual meeting of The Psychonomic Society, November, San Diego, CA.

Doherty, M. E., & Sullivan, J. A. (In press). $\rho = p$. Organizational Behavior and Human Decision Processes.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.). Formal representation of human judgment. NY: Wiley.

Einhorn, H., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. Annual Review of Psychology , **32**, 53-88

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. Psychological Review . 90, 239-260.

Galton, F. (1889) Natural Inheritance. London: MacMillan & Co.

Gorman, M.E. (1986) How the possibility of error affects falsification on a task that models scientific problem solving. British Journal of Psychology, 77, 85-96.

Gorman, M. E. & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory, and a control strategy on Wason's 2-4-6 task. Quarterly Journal of Experimental Psychology, 36A, 629-48.

Gorman, M. E., Gorman, M. E., Latta, R. M. & Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem-solving. British Journal of Psychology, 75, 65-79.

Hammond, K.R. (1986). A theoretically based review of theory and research

in judgment and decision making (Report No. 260). Boulder, Co: University of Colorado.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. Psychological Review. **71**, 438-456.

Hammond, K.R., & Summers, D.A. (1972) Cognitive control. Psychological Bulletin, 79, 58-67.

Himmelfarb, S. (1975). What do you do when the control group doesn't fit into the factorial design? Psychological Bulletin, **82**, 363-368.

Hersh, S. (1986). The target is destroyed. NY: Random House.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. Psychological Review. **71**, 42-60.

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). In D. Kahneman, P. Slovic, & A. Tversky, (Eds.). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Kahneman, D., Slovic, P, & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. London: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Keeley, S. M., & Doherty, M. E. (1971a). A Bayesian prediction of multiple-look identification performance from one-look data: The effect of unequal prior probabilities. Perception and Psychophysics, **10**, 119-122.

Keeley, S. M., & Doherty, M. E. (1971b). Bayesian aggregation of independent successive visual inputs. Journal of Experimental

Psychology, **90**, 300-305.

Kern, L. (1982). The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing. Unpublished Ph.D. dissertation, The Ohio State University.

Kern, L., & Doherty, M. E. (1982). 'Psedodiagnosticity' in an idealized medical problem-solving environment. Journal of Medical Education, **57**, 100-104.

Kirk, R.E. (1982). Experimental Design. Monterey, CA: Brooks/Cole Publishing Company.

Klayman, J., & Ha, Y. W. (1985). Strategy and structure in rule discovery. Paper presented at the Tenth Research Conference on Subjective Probability, Utility and Decision Making, Helsinki, Finland.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94, 211-228.

Mahoney, M. J., & DeMonbreun, B. G. (1978). Psychology of the scientist: An analysis of problem solving bias. Cognitive Therapy and Research , **1**, 229-238.

Markowitz, L. (1983). Effects of self-generated unreliability on predictions. Unpublished Doctoral Dissertation, Bowling Green State University.

Markowitz, L. R. & Mynatt, J. (1982). The effects of data unreliability on the 2-4-6 task. Unpublished manuscript. Bowling Green State University.

Mitroff, I. I. (1974). The subjective side of science. Amsterdam: Elsevier.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 29, 85-95.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. Quarterly Journal of Experimental Psychology, 30, 395-406.

Newell, A., & Simon, H. A. (1972). Human problem solving . Englewood-Cliffs, N. J. : Prentice-Hall

Popper, K. (1959). The logic of scientific discovery. New York: Basic Books.

Schum, D. A. (1977). Contrast effects in inference: On the conditioning of current evidence by prior evidence. Organizational Behavior and Human Performance. **17**, 217-253.

Simon, H. (1957). Models of man: Social and rational. New York: Wiley.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. Journal of Experimental Social Psychology , 22, 93-121.

Slowiaczek, L. M., & Sherman, S. J. (1987, November). Paper presented at the Twenty-eighth Annual Meeting of the Psychonomic Society, Seattle, Washington.

Smedslund, J. (1963). The concept of correlation in adults. Scandinavian Journal of Psychology , **4**, 165-173.

Snyder, M., & Swann, W. B. (1978). Behavioral confirmation in social

interaction: From social perception to social reality. Journal of Experimental Social Psychology, **14**, 148-162.

Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. Journal of Personality and Social Psychology, **43**, 22-34.

Trope, Y., & Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. Journal of Experimental Social Psychology, **19**, 560-576.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. Psychological Review. **71**, 528-530.

Tukey, D. D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. Quarterly Journal of Experimental Psychology, 38A, 5-33.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, **185**, 1124-1131.

Tweney, R. D. (1985). Faraday's discovery of induction: A Cognitive approach. In D. Gooding & F. James (Eds.), Faraday rediscovered (pp. 159-206). London: MacMillan.

Tweney, R. D. & Doherty, M. E. (1983). Rationality and the psychology of inference. Synthese, 57, 139-161.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.) (1981). On scientific thinking. New York: Columbia University Press.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. Quarterly Journal of Experimental Psychology, 32,

109-23.

Von Winterfeldt, D., & Edwards, W. (1986). Decision analysis and behavioral *research* . New York: Cambridge University Press

Walker, B. J. (1985). Instructional effects on strategy choice and solving efficiency in the Wason 2-4-6 rule discovery task. Unpublished master's thesis, Bowling Green State University, Bowling Green, OH.

Walker, B. J. (1986). Variations in problem-solving styles in the Wason 2-4-6 rule discovery task. Paper presented at the first annual Ohio University Inference Conference, Athens, OH

Walker, B.J., Doherty, M.E., & Tweney, R.D. (1987). The effects of feedback error on hypothesis testing. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Walker, B. J. & Tweney, R. D. (1983). Comparison of one and two-rule conditions in the Wason 2-4-6 rule discovery task. Unpublished manuscript, Bowling Green State University, Department of Psychology, Bowling Green, OH.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-40.

Wason, P. C. (1962). Reply to Wetherick. Quarterly Journal of Experimental Psychology, 4, 250.

Wason, P.C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, **23**, 27-281.

Wason, P. C., & Johnson-Laird, P. N. (1972). Psychology of reasoning: Structure and content . Cambridge: Harvard University Press.

Wetherick, N. E. (1962). Eliminative and enumerative behavior in a conceptual task. Quarterly Journal of Experimental Psychology, **4**,

246-249.

York, K., Doherty, M., & Kamouri, J. (1987). The influence of cue unreliability on judgment in a multiple cue probability learning task. Organizational Behavior and Human Decision Processes, **39**, 303-317.